Εξόρυξη Γνώσης από Βάσεις Χωρικών και Χρονικών Δεδομένων

Βασίλειος Μεγαλοοικονόμου, Ph.D.

Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Έλεγχος/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Εισαγωγή

- Η σύγχρονη τεχνολογία καθιστά εφικτή την συλλογή πληθώρας χωρικών και χρονικών δεδομένων → μεγάλες βάσεις δεδομένων
- Πλούτος δεδομένων αλλά πενιχρή πληροφορία
- Ανάγκη για meta-ανάλυση δεδομένων από πολλαπλές μελέτες
- Ανάγκη για robust εργαλεία πληροφορικής για τη στήριξη λήψης αποφάσεων (π.χ., διάγνωση στον τομέα της ιατρικής)
- Στόχος: κατανόηση προτύπων, προσδιορισμός συσχετίσεων, ομαλοτήτων, ανωμαλιών, ...

Εξόρυξη γνώσης?

Αποτελεσματική εξόρυξη ενδιαφέρουσας (μη-τετριμένης, υπονοούμενης, προηγουμένως άγνωστης και πιθανώς χρήσιμης) πληροφορίας ή προτύπων (patterns) από μεγάλες βάσεις δεδομένων

- Πρότυπα (Patterns)?
  - Συσχετίσεις - Associations (π.χ., butter + bread --> milk)
  - Ακολουθίες (π.χ., χρονικά δεδομένα – χρηματιστήριο)
  - Κανόνες για χωρισμό δεδομένων (π.χ., πρόβλημα εντοπισμού θέσης μαγαζιού)
  - …
- Τι είδους πρότυπα είναι ενδιαφέροντα ("**interesting**")?

  Σύμφωνα με:
  - Περιεχόμενο πληροφορίας, confidence and support, p-value, απροσδοκότητα, ικανότητα για εφαρμογή/δράση actionability (utility in decision making)

Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Εκτίμηση/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Το πρόβλημα

Εξόρυξη γνώσης - χωρικά δεδομένα:

Δεδομένης μιας μεγάλης συλλογής από χωρικά δεδομένα, π.χ., 2D ή 3D εικόνες, και άλλα δεδομένα, βρείτε ενδιαφέροντα πράγματα, δηλ.:

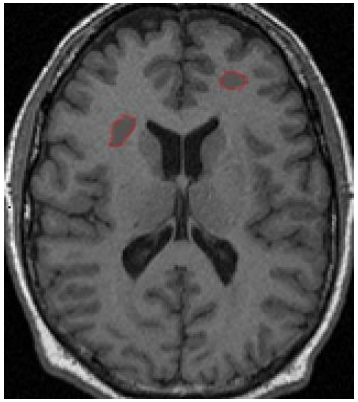
- Συσχετίσεις μεταξύ χωρικών δεδομένων ή μεταξύ χωρικών και μη-χωρικών δεδομένων
- Περιοχές ικανές να διακρίνουν (discriminative areas) ομάδες εικόνων
- Κανόνες/πρότυπα
- Εικόνες παρόμοιες με μια εικόνα query (queries by content)

Χωρικά Δεδομένα - Δυσκολίες

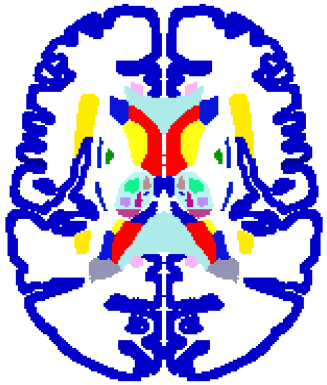
- Πως να εφαρμόσουμε τεχνικές εξόρυξης γνώσης σε εικόνες?
- Ετερογένεια και διακύμανση (variability) δεδομένων
- Προεπεξεργασία δεδομένων (π.χ., κατάτμιση, χωρική κανονικοποίηση, αναπαράσταση)
- Εξερεύνηση της υψηλής συσχέτισης μεταξύ γειτονικών αντικειμένων
- Υψηλή διαστατικότητα δεδομένων (large dimensionality of data)
- Πολυπλοκότητα συσχετίσεων
- Αποδοτική (efficient) διαχείρηση της τοπολογικής πληροφορίας/απόστασης
- Αναπαράσταση χωρικής γνώσης / Spatial Access Methods (SAMs) ώστε να διευκολύνει την αναζήτηση

Υπόβαθρο: Βάσεις Χωρικών Δεδομένων

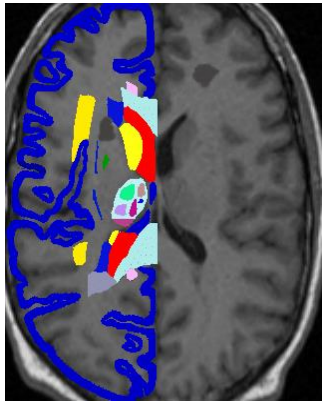
- Εικόνες-δεδομένα: συλλογή από απεικονίσεις σάρωσης σε πολλαπλά επίπεδα
- Εστίαση στην διαχείρηση χωρικών περιοχών ειδικού ενδιαφέροντος (ROIs)
- Προεπεξεργασία:
  - Προσδιορισμός ορίων χωρικών περιοχών ειδικού ενδιαφέροντος
  - Χωρική κανονικοποίηση – (Registration, Spatial Normalization)



Original MRI

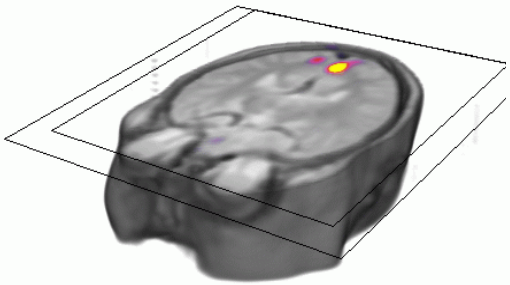


Atlas

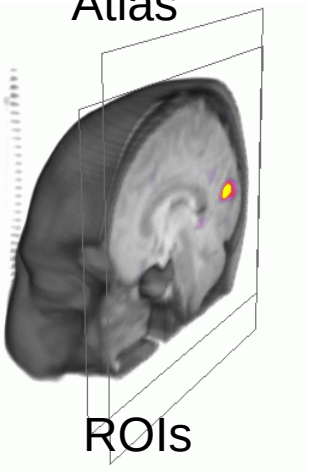


παραμόρφωση MRI

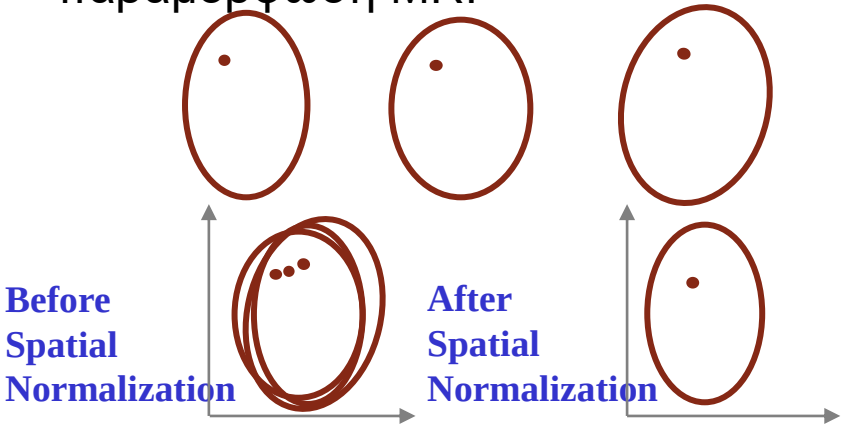
ROIs



ROIs

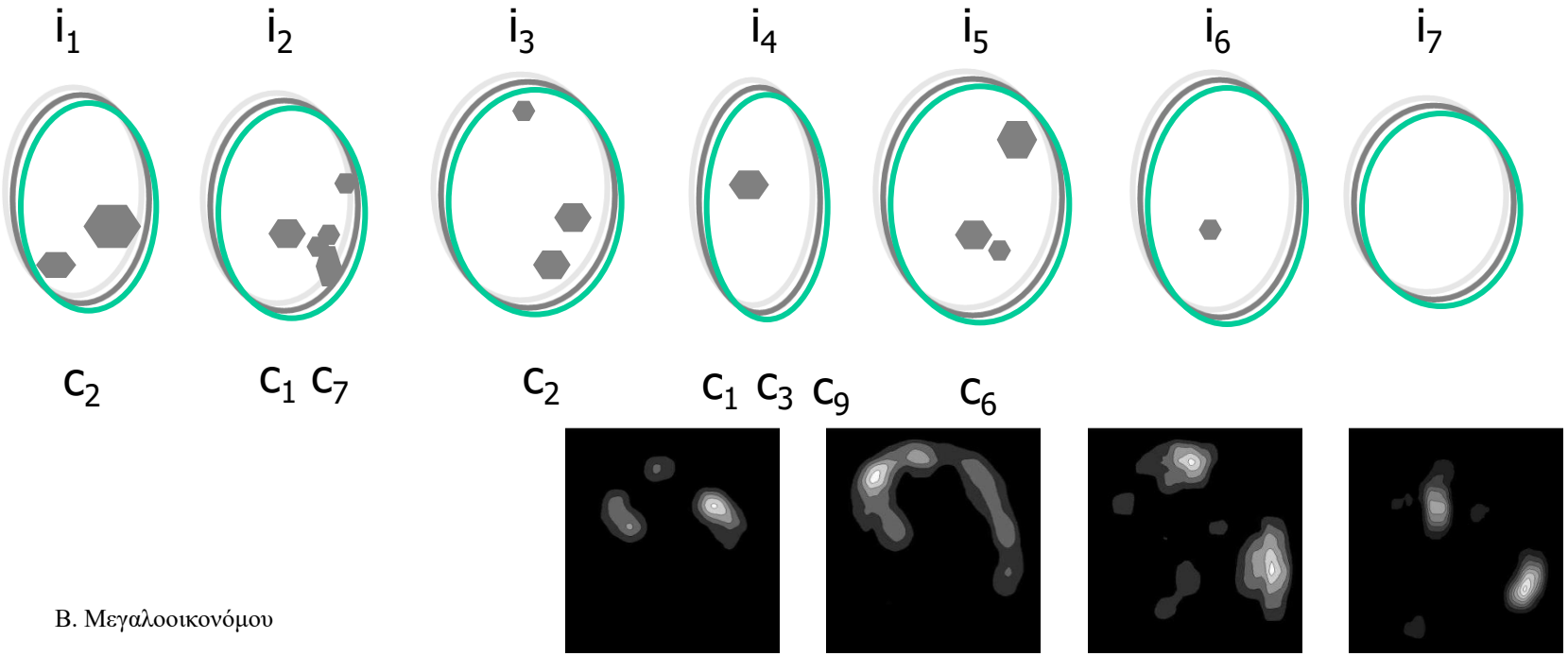


ROIs

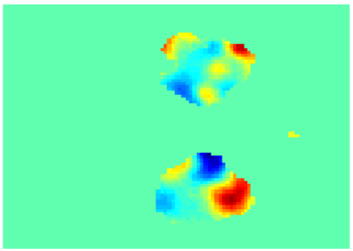


Παράδειγμα: Εξόρυξη Συσχετίσεων

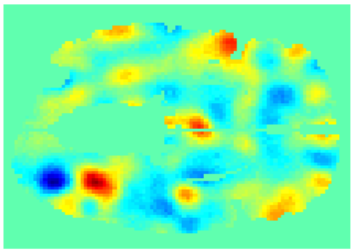
- Ανακάλυψη συσχετίσεων μεταξύ χωρικών και μη-χωρικών δεδομένων (σε μεγάλες Βάσεις Δεδομένων)
  - Εικόνες $\{i_1, i_2, \ldots, i_L\}$
  - Χωρικές περιοχές $\{s_1, s_2, \ldots, s_K\}$
  - Μη-χωρικές μεταβλητές $\{c_1, c_2, \ldots, c_M\}$



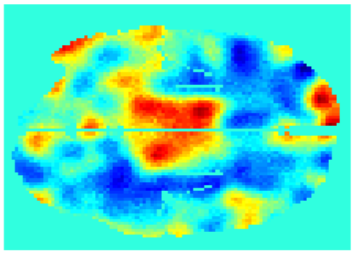
Παράδειγμα: fMRI χάρτες αντίθεσης (contrast maps)



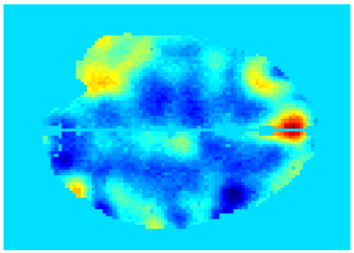
slice 12



slice 24

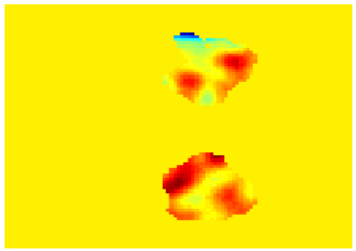


slice 36

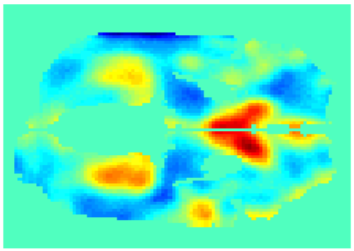


slice 48

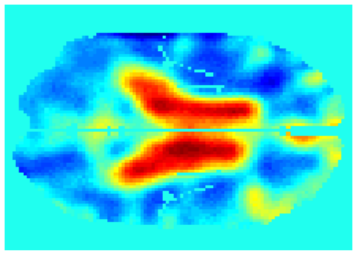
Υγιές δείγμα



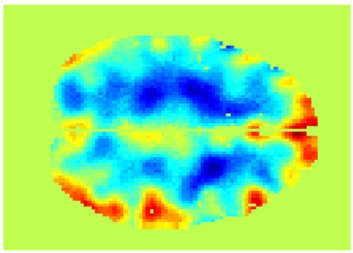
slice 12



slice 24



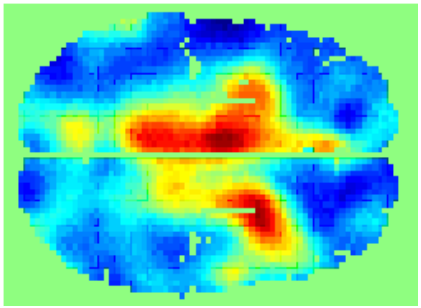
slice 36

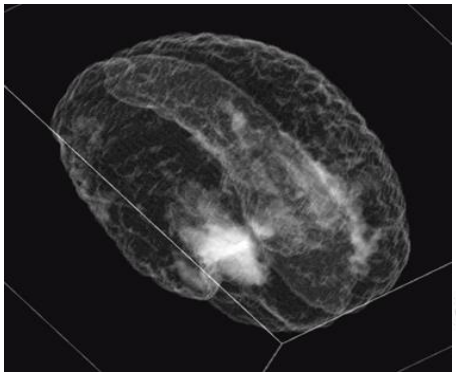


slice 48

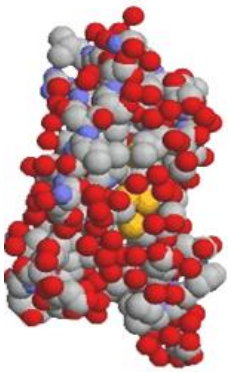
Δείγμα με νόσημα

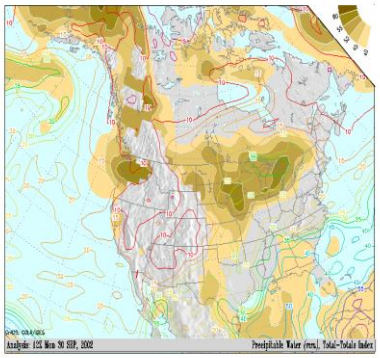
Εφαρμογές

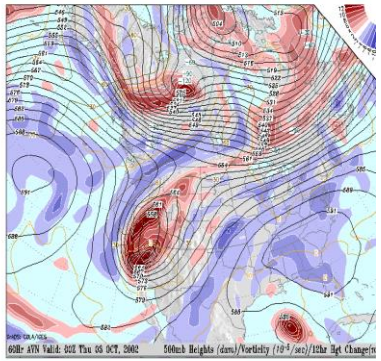


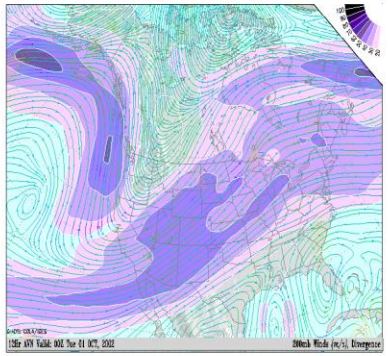


Compression Dilatation









Ανάλυση ιατρικών εικόνων, βιοπληροφορική, γεωγραφία, μετεωρολογία, ...

Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Έλεγχος/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Ανάλυση Χωρικών Δεδομένων voxel-by-voxel

- Μη ύπαρξη μοντέλου για εικόνες
- Η μεταβολή κάθε voxel αναλύεται ανεξάρτητα – παράγεται χάρτης στατιστικής σημαντικότητας
- Η διακριτική ικανότητα υπολογίζεται με στατιστικά τεστς (t-test, ranksum test, F-test, …)
- Statistical Parametric Mapping (SPM)
- Η ισχύς των συσχετίσεων μετριέται με το chi-squared test, Fisher's exact test (πίνακες πιθανών ενδεχομένων για κάθε ζεύγος μεταβλητών)
- Ομαδοποίηση των voxels σύμφωνα με τα αποτελέσματα (λειτουργική σχέση)

[V. Megalooikonomou, C. Davatzikos, E. Herskovits, "Mining Lesion-Deficit Associations in a Brain Image Database", ACM SIGKDD 1999]

Ανάλυση Χωρικών Δεδομένων με Ομαδοποίηση των Voxels

- Ομαδοποίηση των voxels βασισμένη σε άτλαντα
  - Προγενέστερη γνώση αυξάνει την ευαισθησία
  - Ελάττωση δεδομένων: 10^7 voxels ⟶ R περιοχές (δομές)
  - Απεικόνηση κάθε ROI σε τουλάχιστον μια περιοχή
  - Τόσο καλή όσο και ο χάρτης (άτλας) που χρησιμοποιείται
- M μη-χωρικές μεταβλητές, R περιοχές
- Ανάλυση
  - Διακριτές (Categorical) μεταβλητές δομής
    - M x R πίνακες πιθανών ενδεχομένων, Chi-square/Fisher exact test
    - πρόβλημα πολλαπλών συγκρίσεων
    - log-linear analysis, multivariate Bayesian
  - Συνεχείς (Continuous) μεταβλητές δομής
    - Logistic regression, Mann-Whitney

[V. Megalooikonomou, C. Davatzikos, E. Herskovits, "Mining Lesion-Deficit Associations in a Brain Image Database", ACM SIGKDD 1999]

Μέθοδοι: Στατιστικές: Chi-square

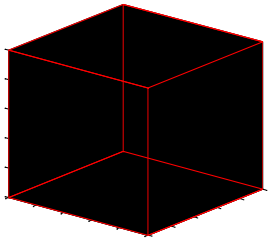
- 2 x 2 πίνακες πιθανών ενδεχομένων για categorical μεταβλητές

		lesion in s_i		
		No	Yes	
function f_k	No	n_{11}	n_{12}	$n_{1.}$
	Yes	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	n

- Pearson chi-square

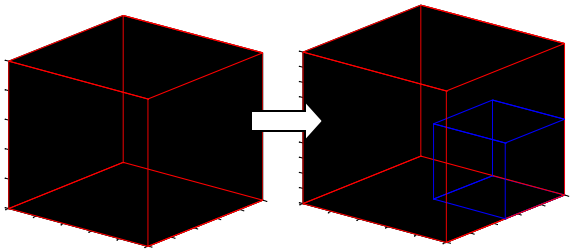
$$\chi^2 = \sum \frac{(n_{ij} - m_{ij})^2}{m_{ij}}, \quad m_{ij} = \frac{n_{i.} n_{.j}}{n}$$

Δυναμικός Αναδρομικός Διαμερισμός (Dynamic Recursive Partitioning)



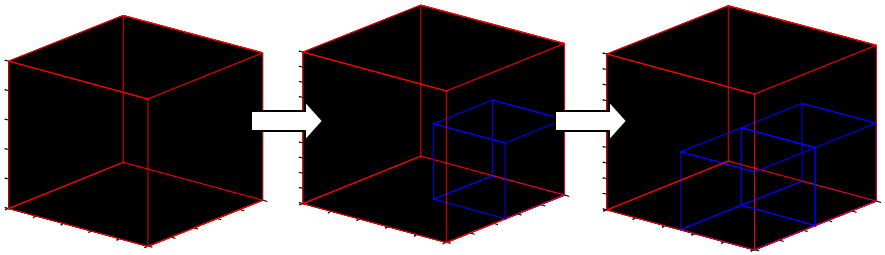
- Δυναμικός αναδρομικός διαμερισμός ενός τρισδιάστατου χώρου σε υποπεριοχές

Δυναμικός Αναδρομικός Διαμερισμός (Dynamic Recursive Partitioning)



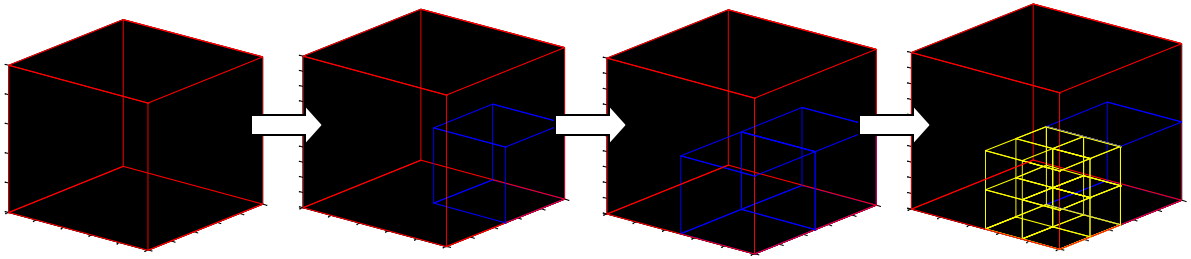
- Δυναμικός αναδρομικός διαμερισμός ενός τρισδιάστατου χώρου σε υποπεριοχές
- Κριτήρια για τον διαμερισμό:
  - διακριτική ισχύ (discriminative power) των χαρακτηριστικών της υποπεριοχής (hyper-rectangle) και
  - μέγεθος της υποπεριοχής

Δυναμικός Αναδρομικός Διαμερισμός (Dynamic Recursive Partitioning)



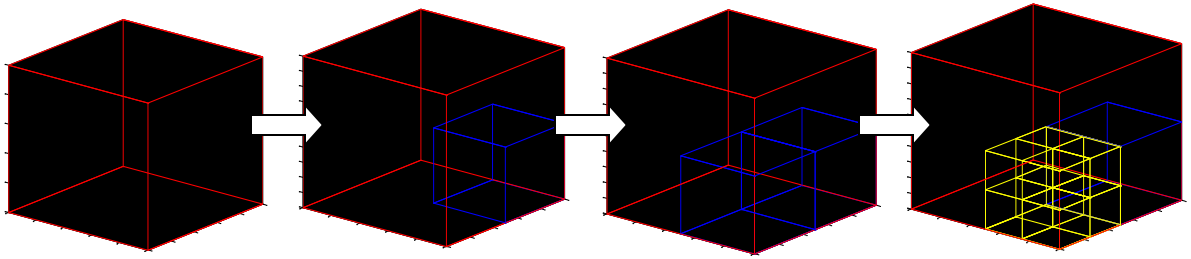
- Δυναμικός αναδρομικός διαμερισμός ενός τρισδιάστατου χώρου σε υποπεριοχές
- Κριτήρια για τον διαμερισμό:
  - διακριτική ισχύ (discriminative power) των χαρακτηριστικών της υποπεριοχής (hyper-rectangle) και
  - μέγεθος της υποπεριοχής

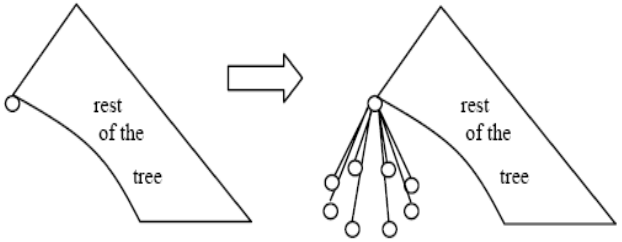
Δυναμικός Αναδρομικός Διαμερισμός (Dynamic Recursive Partitioning)



- Δυναμικός αναδρομικός διαμερισμός ενός τρισδιάστατου χώρου σε υποπεριοχές
- Κριτήρια για τον διαμερισμό:
  - διακριτική ισχύ (discriminative power) των χαρακτηριστικών της υποπεριοχής (hyper-rectangle) και
  - μέγεθος της υποπεριοχής

Δυναμικός Αναδρομικός Διαμερισμός (Dynamic Recursive Partitioning)





- Δυναμικός αναδρομικός διαμερισμός ενός τρισδιάστατου χώρου σε υποπεριοχές
- Κριτήρια για τον διαμερισμό:
  - διακριτική ισχύ (discriminative power) των χαρακτηριστικών της υποπεριοχής (hyper-rectangle) και
  - μέγεθος της υποπεριοχής
- Εξαγωγή χαρακτηριστικών από περιοχές με διακριτική ισχύ για ταξινόμηση σε κλάσεις
- Ελάττωση του προβλήματος πολλαπλών συγκρίσεων (multiple comparison problem) → (# τέστς = # υποδιαιρέσεων < # voxels)
- τεστς downward closed

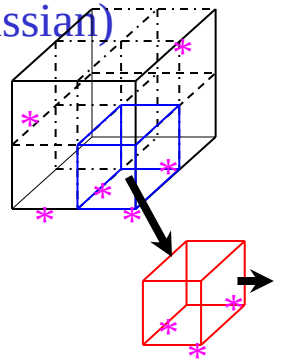
[V. Megalooikonomou, D. Pokrajac, A. Lazarevic, and Z. Obradovic, “Effective classification of 3-D image data using partitioning methods”, *SPIE Conference on Visualization and Data Analysis*, 2002]

[V. Megalooikonomou, D. Kontos, D. Pokrajac, A. Lazarevic, Z. Obradovic, " An adaptive partitioning approach for mining discriminant regions in 3D image data", *Journal of Intelligent Information Systems*, (in review)]

Άλλες μέθοδοι για Ταξινόμηση Χωρικών Δεδομένων

Διάκριση μεταξύ κατανομών:

- Με χρήση αποστάσεων κατανομών πιθανότητας:
  - Mahalanobis distance
  - Kullback-Leibler divergence (parametric, non-parametric)
- Με χρήση Maximum Likelihood (ότι η πιθανοτική κατανομή του νέου αντικειμένου είναι η ίδια με κάποια από τις υπάρχουσες):
  - Εκτίμηση πυκνότητας πιθανότητας (probability densities) και υπολογισμός likelihood
    - Μέθοδος EM (Expectation-Maximization) για μοντελοποίηση χρησιμοποιώντας κάποια βασική συνάρτηση - base function (Gaussian)
- Στατικός διαμερισμός:
  - Ελάττωση των μεταβλητών σε σχέση με την ανάλυση voxel-by-voxel
  - Ο χώρος υποδιαιρείται σε 3D υπερ-πολύγωνα (discretization step)
  - Μεταβλητές: χαρακτηριστικά των voxels μέσα στα υπερ-πολύγωνα
  - Σταδιακή αύξηση ανάλυσης (discretization)



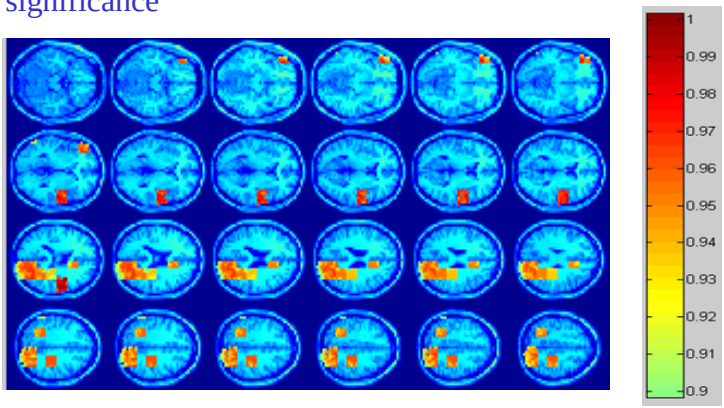
D. Pokrajac, V. Megalooikonomou, A. Lazarevic, D. Kontos, Z. Obradovic, Artificial Intelligence in Medicine, Vol. 33, No. 3, pp. 261-280, Mar. 2005.

Πειραματικά Αποτελέσματα

Method				Classification Accuracy (%)		
	Criterion	Threshold	Tree depth	Controls	Patients	Total
DRP	correlation	0.4	3	82	93	88
	t-test	0.05	3	89	100	94
		0.05	4	84	100	92
		0.01	4	87	100	93
	ranksum	0.05	3	87	100	93
		0.05	4	80	100	90
		0.01	4	87	96	91
Maximum Likelihood / EM				77	67	72
Maximum Likelihood / k-means				77	83	80
Kullback-Leibler / EM				79	57	68
Kullback-Leibler / k-means				77	66	71

Συγκριτικά αποτελέσματα ταξινόμησης χρησιμοποιώντας Maximum Likelihood και EM ή k-means για εκτίμηση της κατανομής

Areas discovered by DRP with t-test: significance threshold=0.05, maximum tree depth=3. Colorbar shows significance



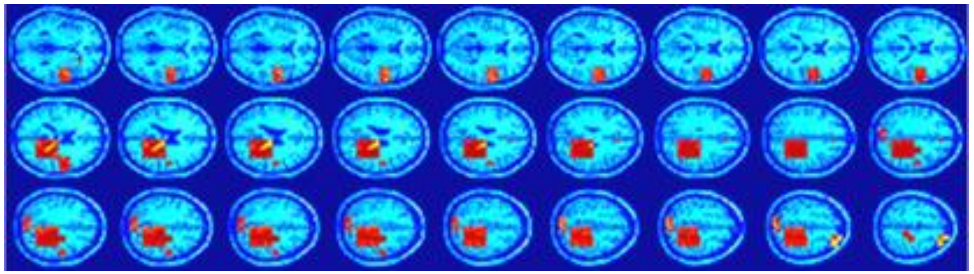
Σύγκριση του αριθμού των τεστς που έχουν πραγματοποιηθεί

		Number of tests	
Thresh	**Depth**	**DRP**	**Voxel Wise**
0.05	3	569	201774
0.05	4	4425	201774
0.01	4	4665	201774

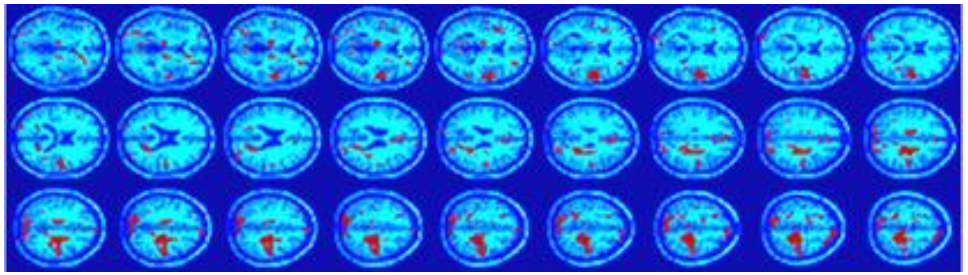
[D. Kontos, V. Megalooikonomou, D. Pokrajac, A. Lazarevic, Z. Obradovic, O. B. Boyko, J. Ford, F. Makedon, A. J. Saykin, MICCAI 2004]

Πειραματικά Αποτελέσματα

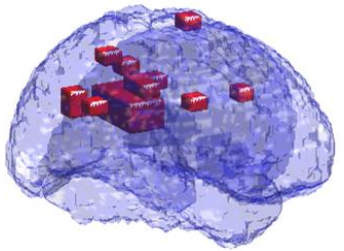
(a)



(b)



Discriminative υπο-περιοχές που εντοπίστηκαν εφαρμόζοντας (a) DRP και (b) ανάλυση voxel-wise με ranksum test και κατώφλι σπουδαιότητας (significance) 0.05 σε κλινικά fMRI δεδομένα



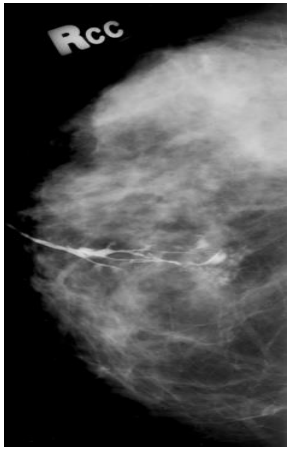
Επίδραση αυτής της εργασίας:

- Βοηθά στην ερμηνεία των εικόνων (π.χ., διευκολύνει την κλινική διάγνωση)
- Καθιστά δυνατή την ενοποίηση, διαχειρισμό και ανάλυση μεγάλων όγκων εικόνων και άλλων τύπων δεδομένων
- Παρέχει νέα γνώση ως προς την συσχέτηση δομής και λειτουργίας

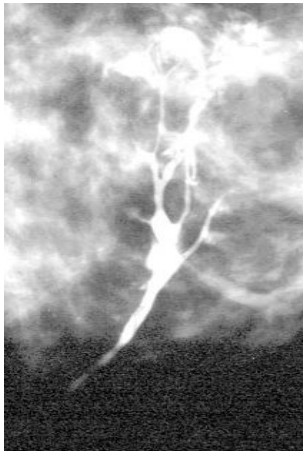
Διαχείρηση Χωρικών Περιοχών

- Λεπτομερής ποσοτικός χαρακτηρισμός και ταξινόμηση (classification) των περιοχών ειδικού ενδιαφέροντος (ROIs) – π.χ., επέκταση ανάλυσης Sholl με ομόκεντρα κέλυφη
- Εξαγωγή και αναπαράσταση του πληροφοριακού περιεχομένου (information content) των εικόνων
- Ανάπτυξη ενός γενικού και ενοποιημένου πλαισίου εργασίας για διαχείρηση περιοχών (ROIs)
- Πιθανοτικά μοντέλα για παραγωγή προσομοιωμένων περιοχών
- Αναζήτηση ομοιότητας – similarity searches

Ποσοτικός Χαρακτηρισμός Δενδροειδών δομών και Ταξινόμηση



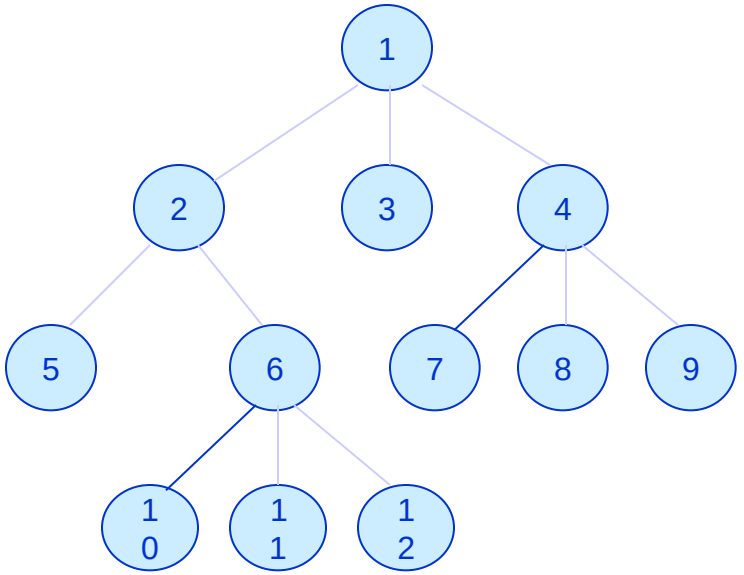
(a)



(b)

(c)

Ένα δίκτυο αγωγών στο μαστό: (a) γαλακτόγραμμα με a contrast-enhanced δίκτυο αγωγών, (b) μέρος του γαλακτογράμματος που δείχνει μεγαλύτερο το δίκτυο αγωγών, (c) το δίκτυο



{1 2 2 6 6 6 1 1 4 4 4 }

Prufer

- Προεπεξεργασία (προσδιορισμός ορίων χωρικών περιοχών, σκελετοποίηση, κανονικοποίηση (ισόμορφα δένδρα)
- Labeling και αναπαράσταση δέντρων με σειρές χαρακτήρων – κωδικοποίηση Prüfer

[V. Megalooikonomou, D. Kontos, J. Danglemaier, A. Javadi, P. A. Bakic, A.D.A. Maidment, Proceedings of the SPIE Conference on Medical Imaging, 2006.]

Ποσοτικός Χαρακτηρισμός Δενδροειδών δομών και Ταξινόμηση

- … αναπαράσταση δέντρων με σειρές χαρακτήρων
- Χρήση τεχνικών tf-idf εξόρυξης γνώσης από κείμενα για ανάθεση βάρους σπουδαιότητας σε κάθε όρο- label
  - Το βάρος w_{ij} του όρου i στη σειρά j προσδιορίζεται ως εξής:

    $$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N/ df_i)$$

    όπου f_{ij} είναι η συχνότητα εμφάνισης του όρου i στη σειρά j,

    $tf_{ij} = f_{ij} / max\{f_{ij}\}$

    df_i = αριθμός των σειρών που περιλαμβάνουν τον όρο i,

    idf_i = αντίστροφο της df_i, = $\log_2 (N/ df_i)$ και

    N: ο συνολικός αριθμός σειρών
  - Η κάθε σειρά αναπαρίστανται ως ένα t-dimensional διάνυσμα:

    $d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$, όπου t = |vocabulary|=διάσταση
  - Δύο σειρές είναι παρεμφερείς με βάση το cosine similarity measure των διανυσμάτων που υπολογίζεται ως εξής:

    $$\text{CosSim}(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{\left|\vec{d_j}\right| \cdot \left|\vec{q}\right|} = \frac{\sum_{i=1}^{t} (w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$$

Ποσοτικός Χαρακτηρισμός Δενδροειδών δομών και Ταξινόμηση

- Similarity searches:
  - Υπολογίζουμε το pairwise cosine distance matrix για όλα τα *tf-idf* διανύσματα.
  - Χρησιμοποιούμε κάθε δένδρο (δηλ. *tf-idf διάνυσμα*) σαν query και βρίσκουμε τα *k* πιο όμοια δέντρα με βάση το cosine distance matrix.
  - Precision: το ποσοστό των σχετικών δένδρων (relevant trees) μεταξύ αυτών που βρέθηκαν – μέσος όρος για όλα τα similarity queries που κάναμε (σχετικά: τα δένδρα που ανήκουν στην ίδια ομάδα με το query tree (NF vs. RF)).

Prufer Encoding			
k	Precision		
	NF	RF	Total
1	100%	100 %	**100%**
2	90.00 %	70.83 %	**79.55 %**
3	80.00 %	66.67 %	**72.73 %**
4	72.50 %	64.58 %	68.18 %
5	68.00 %	65.00 %	66.36 %

Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Έλεγχος/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Έλεγχος/Αξιολόγηση Τεχνικών Εξόρυξης Γνώσης

- Γιατί να ελέγξουμε την διαδικασία εξόρυξης γνώσης?
  - μη γνώση της πραγματικότητας
  - αντικειμενική αξιολόγηση των προσόντων και περιορισμών των μεθόδων
  - υπολογισμός μεγέθους δείγματος
  - επίδραση μεθόδων προεπεξεργασίας δεδομένων
- Κατασκευή χωρικού προσομοιωτή και εκτέλεση Monte Carlo ανάλυσης (simulations)
  - Παραγωγή μεγάλου αριθμού δειγμάτων
  - Χρήση πιθανοτικών κατανομών για μοντελοποίηση:
    - χωρικών περιοχών ειδικού ενδιαφέροντος (αριθμός, μέγεθος, σχήμα, και θέση)
    - μη-χωρικών μεταβλητών
    - συσχετίσεων (ισχύς, αριθμός, δομή)
    - σφάλματος χωρικής κανονικοποίησης και κατάτμησης (segmentation) (των χωρικών περιοχών)

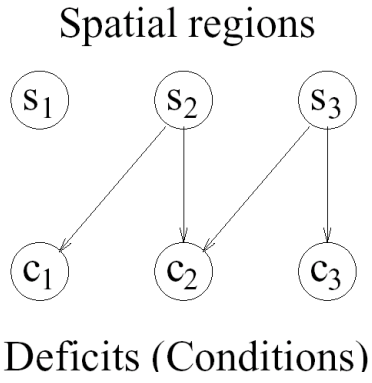
    …σύμφωνα με αληθινά δεδομένα/μελέτες
- Ανακάλυψη συσχετίσεων με εφαρμογή τεχνικών εξόριξης γνώσης
- Σύγκριση των αποκαλυφθέντων συσχετίσεων με τις πραγματικές

Ο Χωρικός Προσομοιωτής

- Χρήση μοντέλου συσχετίσεων μεταξύ χωρικών περιοχών και άλλων μη-χωρικών μεταβλητών (που παρέχονται σαν input ή παράγονται τυχαία χρησιμοποιώντας διάφορες παραμέτρους)
- Χρήση ενός σετ δεδομένων *D* (real study) για εξαγωγή παραμέτρων για:
  - χωρικές και μη-χωρικές μεταβλητές περιοχές, priors, σφάλμα χωρικής κανονικοποίησης (spatial normalization error)
- Παραγωγή προσομοιωμένων δειγμάτων (samples)
  - Για κάθε δείγμα p:
    - παραγωγή προσομοιωμένων χωρικών περιοχών
    - αλλαγή θέσης και παραμόρφωση των περιοχών (modeling spatial normalization error)
    - αναγνώριση περιοχών που είναι ασυνήθιστες (ανώμαλες)
    - παραγωγή προσομοιωμένων μη-χωρικών μεταβλητών με βάση τις τεχνητές συσχετίσεις και τις ανώμαλες περιοχές

Ο χωρικός προσομοιωτής: Μοντελοποίηση συσχετίσεων με χρήση Bayesian Networks

- Μοντέλο Bayesian Network (BN):
  - directed acyclic graph (DAG)
  - απεικονίζει την joint pdf του πεδίου, πληροφορία αιτίων/ανεξαρτησίας
  - κάθε κόμβος είναι μια μεταβλητή (χωρική περιοχή ή μη-χωρική μεταβλητή (δυσ)λειτουργία/κατάσταση)
  - πίνακες δείχνουν conditional πιθανότητες
  - Αρχή *conditional independence*
- Πλεονεκτήματα:
  - μαθηματικά έγκυρη (θεωρία πιθανοτήτων)
  - εκφράζεται παραστατικά
  - έχει δομική σημασία (μπορούν να απεικονίσουν παραγωγικές διαδικασίες (generative processes))
  - πλήρης (οποιαδήποτε παραγωγική διαδικασία)

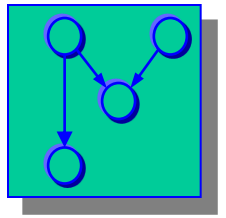
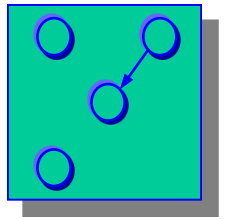
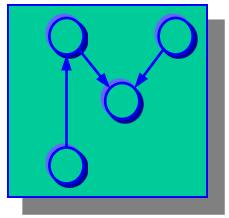
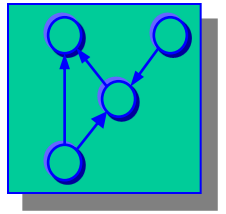


node	conditional-probability table
s_1	$p(s_1) = 0.6$
s_2	$p(s_2) = 0.8$
s_3	$p(s_3) = 0.8$
c_1	$p(c_1\|s_2) = 1.0,\ p(c_1\|\overline{s_2}) = 0.6$
c_2	$p(c_2\|s_2, s_3) = 0.4,\ p(c_2\|s_2, \overline{s_3}) = 0.9,$
	$p(c_2\|\overline{s_2}, s_3) = 0.2,\ p(c_2\|\overline{s_2}, \overline{s_3}) = 0.9$
c_3	$p(c_3\|s_3) = 0.3,\ p(c_3\|\overline{s_3}) = 0.9$

•V. Megalooikonomou, C. Davatzikos, E. Herskovits, "A Simulator for Evaluation of Methods for the Detection of Lesion-Deficit Associations", ***Human Brain Mapping***, Vol. 10, No. 2, pp.61-73, 2000.

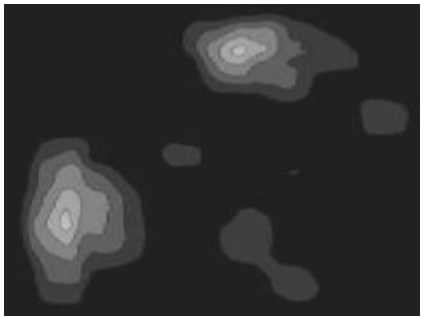
Ο χωρικός προσομοιωτής: Μοντελοποίηση συσχετίσεων με χρήση Bayesian Networks

- Δημιουργία τυχαίων μοντέλων Bayesian Networks χρησιμοποιώντας τις παρακάτω παραμέτρους:
  - αριθμό κόμβων χωρικών περιοχών και μη χωρικών μεταβλητών
  - αριθμό συσχετίσεων (associations)
  - prior πιθανότητα κάθε περιοχής να είναι ενδιαφέρουσα
  - conditional πιθανότητες (δυσ)λειτουργιών (π.χ., ισχύς των συσχετίσεων)
  - βαθμός (degree) συσχέτισης κόμβων, ανώτατο όριο βαθμού του δικτύου
- Χρήση noisy-OR model για probabilistic disjunctive αλληλεπιδράσεις μεταξύ αιτίων ενός αποτελέσματος (leak prob.)
- Υπολογισμός prior πιθανότητες για τις χωρικές περιοχές
  - υπολογισμός ποσοστού επί τοις εκατό των δειγμάτων με ανωμαλία σε κάθε χωρική περιοχή (χρήση sigmoid ή step function)

Ο χωρικός προσομοιωτής: Παραγωγή προσομοιωμένων περιοχών (ROIs)

- Δημιουργία προσομοιωμένων περιοχών χρησιμοποιώντας στατιστικά δεδομένα συλλεγμένα από ένα σετ δεδομένων *D* για:
  - αριθμό των περιοχών ειδικού ενδιαφέροντος / δείγμα
  - μέγεθος κάθε περιοχής
  - σχήμα κάθε περιοχής
  - χωρική πιθανοτική κατανομή του κέντρου μάζας κάθε περιοχής
    - εξομάλυνση (κάνοντας χρήση ενός σφαιρικού παραθύρου), padding, κανονικοποίηση (normalization)

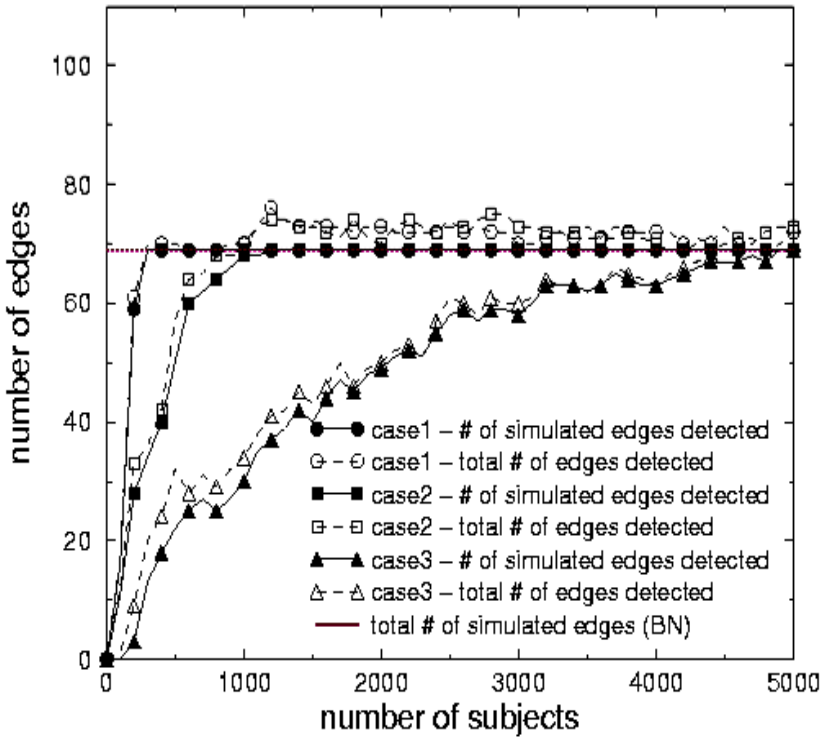


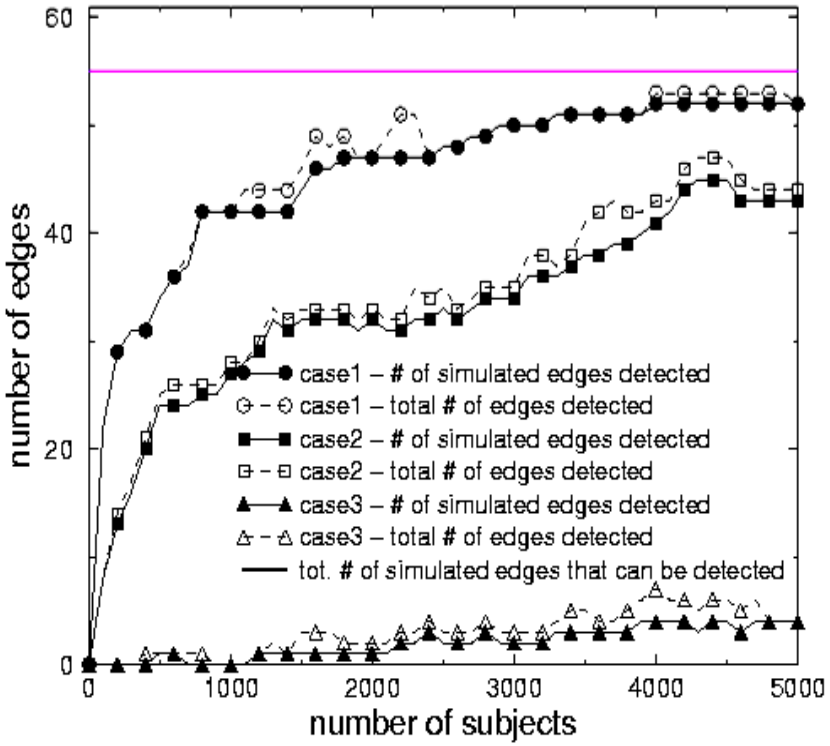
		6			
		5	4	6	
	4	1	3	5	
	5	2	4	8	
		7	5		

- Δημιουργία των σχημάτων των περιοχών χρησιμοποιώντας ένα μοντέλο ανάπτυξης (growth model) βασισμένο σε αυτό του Eden, 1961

V. Megalooikonomou, "Evaluating the performance of association mining methods in 3-D medical image databases", *SIAM International Conference on Data Mining*, 2002.

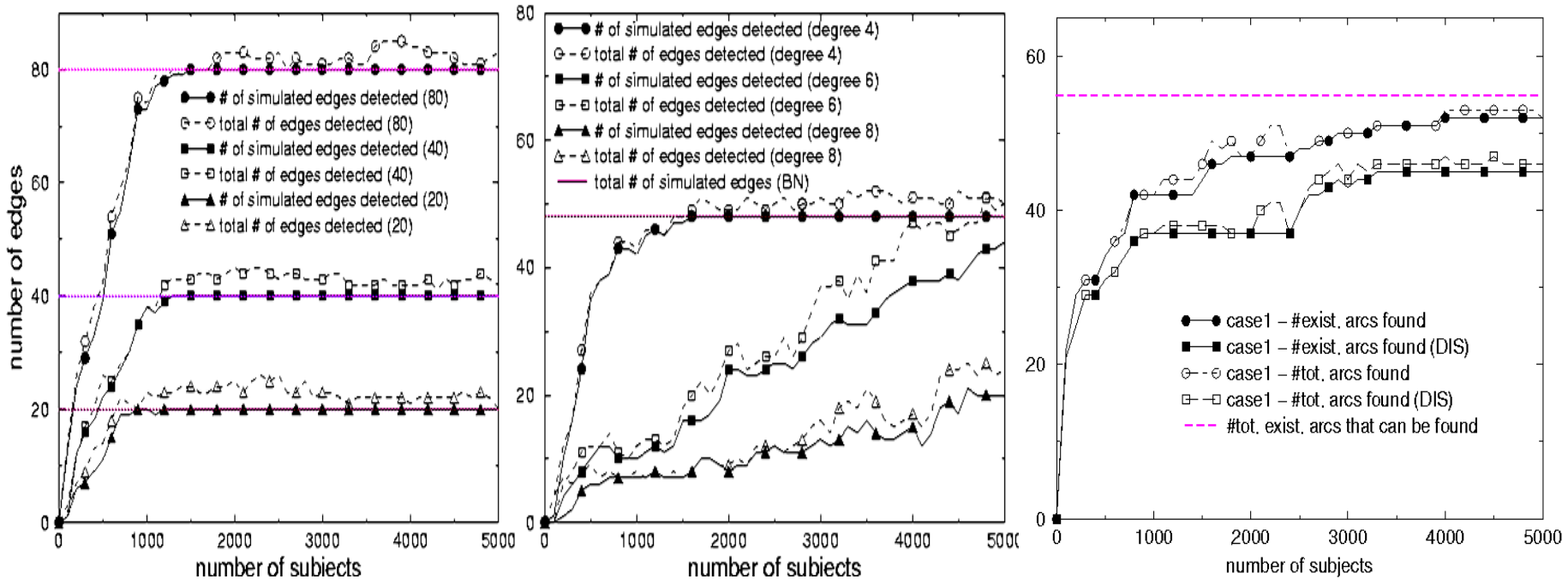
Αποτελέσματα: Χωρικός Προσομοιωτής





- Ν είναι αντιστρόφως ανάλογο της μικρότερης prior/conditional πιθανότητας

Αποτελέσματα: Χωρικός Προσομοιωτής



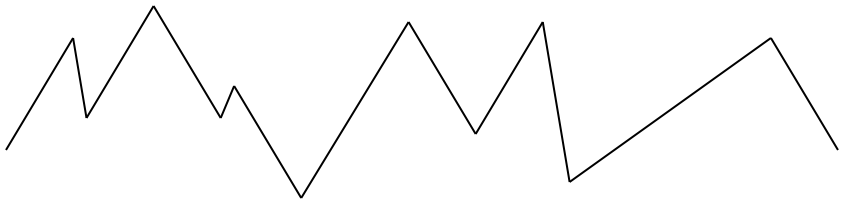
- Ο βαθμός (degree) συσχέτισης επηρεάζει περισσότερο την απόδοση του συστήματος από τον αριθμό των υπάρχοντων συσχετίσεων
- Κατά μέσον όρο 87% των συσχετίσεων βρέθηκαν σε ευθυγραμμισμένες εικόνες σε σύγκριση με τέλεια ευθυγράμμιση (perfect registration)

Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Εκτίμηση/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Ανάλυση Χρονικών Σειρών

Χρονική σειρά: Μία ακολουθία (διατεταγμένη συλλογή) πραγματικών τιμών : $X = x_1, x_2, \ldots, x_n$



- Πολλές εφαρμογές ...
- Δυσκολίες:
  - Υψηλή διαστατικότητα δεδομένων (high dimensionality)
  - Μεγάλος αριθμός σειρών
  - Ορισμός συνάρτησης ομοιότητας (similarity)
- Ανάλυση ομοιότητας (π.χ., ανεύρεση μετοχών που μοιάζουν με αυτήν της IBM
- Στόχοι: υψηλή ακρίβεια, γρήγορες τεχνικές για αναζήτηση ομοιότητας μεταξύ χρονικών σειρών και για ανακάλυψη ενδιαφέροντων προτύπων
- Εφαρμογές: ομαδοποίηση (clustering), ταξινόμηση (classification), αναζήτηση ομοιότητας (similarity searches), δημιουργία περίληψης (summarization)

Τεχνικές Μείωσης Διαστατικότητας (Dimensionality)

- DFT: Discrete Fourier Transform
- DWT: Discrete Wavelet Transform
- SVD: Singular Value Decomposition
- APCA: Adaptive Piecewise Constant Approximation
- PAA: Piecewise Aggregate Approximation
- SAX: Symbolic Aggregate approXimation
- …

Αποστάσεις ομοιότητας χρονικών σειρών (similarity distances)

- Ευκλείδειος Απόσταση:

  η πιο διαδεδομένη αλλά ευπαθής σε μεταβολές (shifts)

- Δυναμική παραμόρφωση χρόνου (Dynamic Time Warping):

  βελτιώνει την ακρίβεια αλλά είναι αργή: $O(n^2)$

- Envelope-based DTW:

  πιο γρήγορη: O(n)

Μια πιο διαισθητική ιδέα:

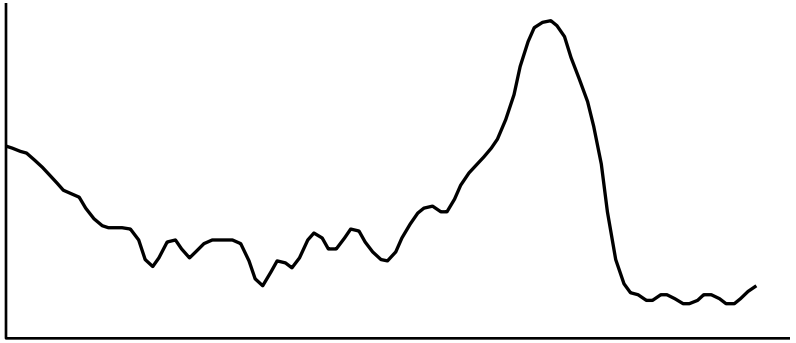
δύο σειρές πρέπει να θεωρούνται όμοιες αν έχουν αρκετά μη-επικαλυπτόμενα χρονικά διατεταγμένα ζεύγη υποσειρών που είναι όμοια (Agrawal et al. VLDB, 1995)

Κύρια σημεία

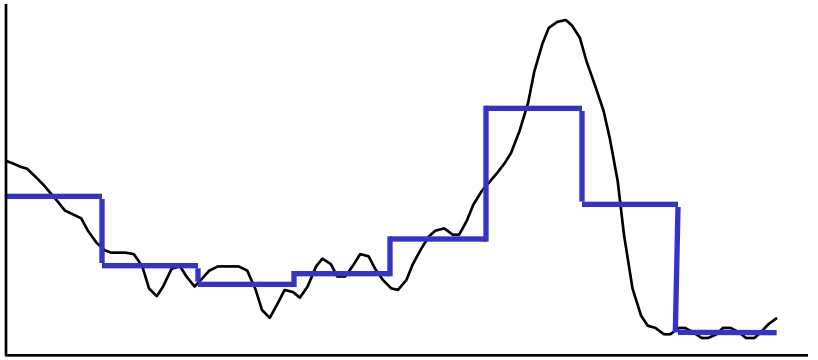
- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Έλεγχος/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Διαμελισμός – Προσέγγιση Τμημάτων με Σταθερές (Piecewise Constant Approximations)

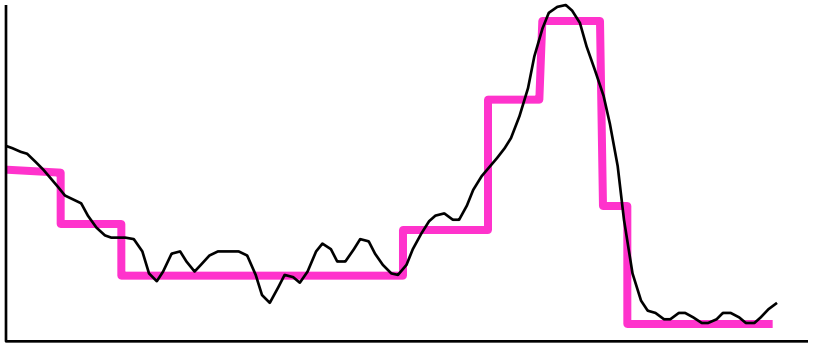
Αρχική σειρά
(n σημεία)



Piecewise constant approximation (PCA) ή Piecewise Aggregate Approximation (PAA), [Yi and Faloutsos ’00, Keogh et al, ’00]
(n' τμήματα)



Προσαρμόσιμη PCA - Adaptive Piecewise Constant Approximation (APCA), [Keogh et al., ’01]
(n'' τμήματα)



Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Έλεγχος/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Multiresolution Vector Quantized (MVQ) προσέγγιση

Διαμελίζει μια σειρά σε τμήματα ίσου μήκους και χρησιμοποιεί VQ για αναπαράσταση κάθε τμήματος – η σειρά αναπαριστάνεται με τη συχνότητα ενφάνισης υποσειρών κλειδίων

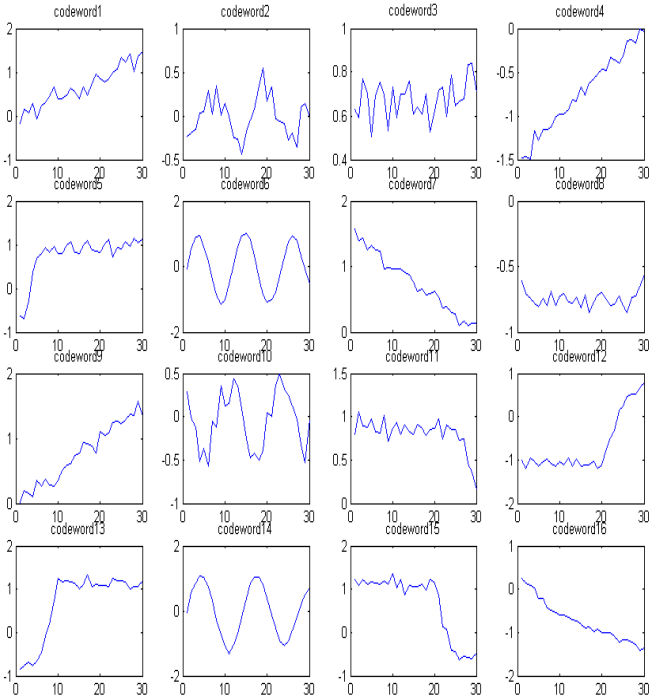
1) Χρησιμοποιεί ένα 'λεξικό' υποσειρών (βιβλίο κώδικα - codebook) – χρησιμοποιείται κάποιο training

2) Χρησιμοποιεί πολλαπλή ανάλυση (multiple resolutions) – διατηρεί τοπική και γενική πληροφορία

3) Αντίθετα από τα wavelets αγνοεί μερικώς την σειρά (διάταξη) των κωδικών ('codewords')

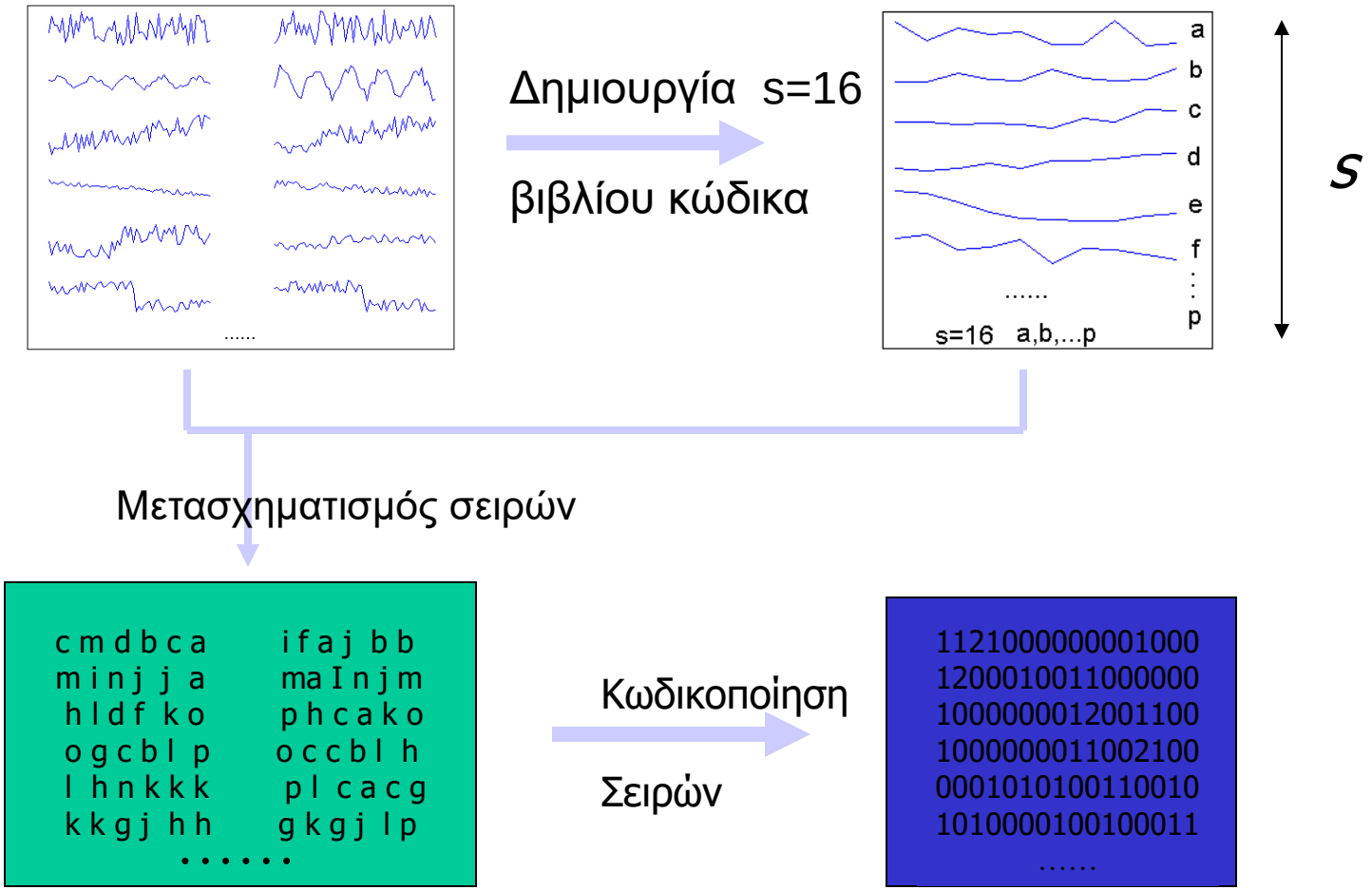
3) Εκμεταλλεύεται προηγούμενη γνώση (prior knowledge) σχετικά με τα δεδομένα

4) Χρησιμοποιεί μια νέα συνάρτηση απόστασης (distance function)



[V. Megalooikonomou, Q. Wang, G. Li, C. Faloutsos, ICDE 2005]

Μεθοδολογία



Μεθοδολογία

-Δημιουργία ‘λεξικού’

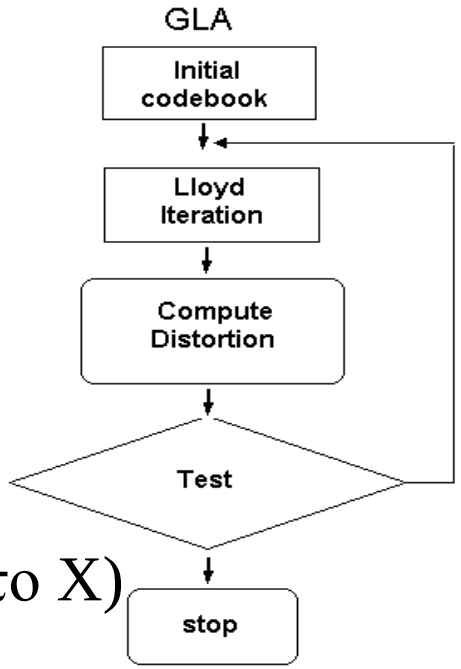
Q: Πως?

A: Χρήση Vector Quantization, πιο συγκεκριμένα, χρήση του Generalized Lloyd Algorithm (GLA)

Πρότυπα που εμφανίζονται συχνά σε υποσειρές

Output:

Βιβλίο κώδικα με *s* κωδικούς (codewords)



-Αναπαράσταση χρονικών σειρών

$X = x_1, x_2, \ldots, x_n$

Κωδικοποιείται με την νέα αναπαράσταση

$f = (f_1, f_2, \ldots, f_s)$

(f_i : η συχνότητα εμφάνισης του i th κωδικού στο X)

Μεθοδολογία

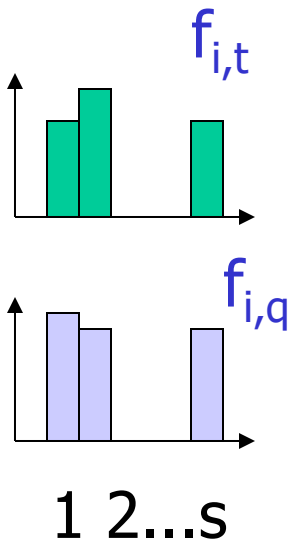
Νέα συνάρτηση απόστασης:

Το μοντέλο ιστογράμματος (HM) χρησιμοποιείται για να υπολογιστεί η ομοιότητα μεταξύ των αναπαραστάσεων δυο χρονικών σειρών σε κάθε επίπεδο ανάλυσης (resolution level):

$$S_{HM}(q,t) = \frac{1}{1 + dis(q,t)}$$

όπου

$$dis(q,t) = \sum_{i=1}^{s} \frac{\left| f_{i,t} - f_{i,q} \right|}{1 + f_{i,t} + f_{i,q}}$$



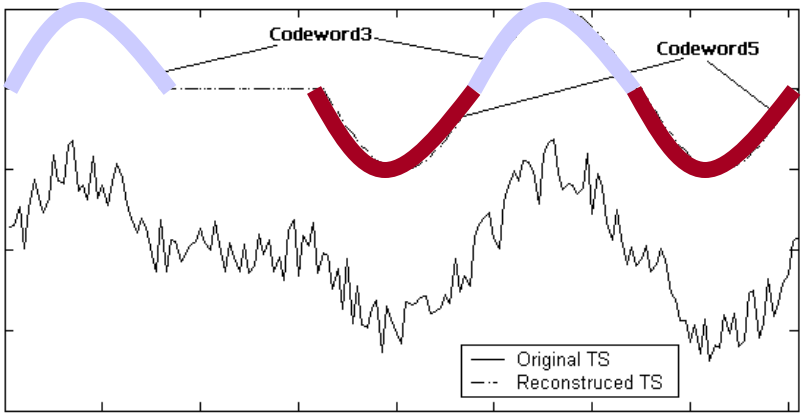
Μεθοδολογία

Δημιουργία περίληψης χρονικών σειρών (Time series summarization):

• Πληροφορία υψηλού επιπέδου (πρότυπα που εμφανίζονται συχνά) είναι πιο χρήσιμη

• Η καινούργια αναπαράσταση παρέχει αυτό το είδος της πληροφορίας

Οι κωδικοί (patterns) 3 & 5 εμφανίζονται 2 φορές



Μεθοδολογία

Προβλήματα με την κωδικοποίηση που χρησιμοποιεί συχνότητα εμφάνισης :

- Δύσκολο να καθοριστεί το σωστό επίπεδο ανάλυσης (resolution - codeword length)
- Μπορεί να χάσει γενική πληροφορία

Μεθοδολογία

Λύση: Χρήση πολλαπλών επιπέδων ανάλυσης (multiple resolutions):

- Δύσκολο να καθοριστεί το σωστό επίπεδο ανάλυσης (resolution - codeword length)
- Μπορεί να χάσει γενική πληροφορία

Μεθοδολογία

Η προτεινόμενη συνάρτηση απόστασης:

Άθροισμα ομοιοτήτων με βάρη, σε όλα τα επίπεδα ανάλυσης (resolution levels) – μοντέλο ιεραρχικού ιστογράμματος

$$S_{HHM}(q, d_j) = \sum_{i=1}^{c} w_i * S_{HMi}(q, d_j)$$

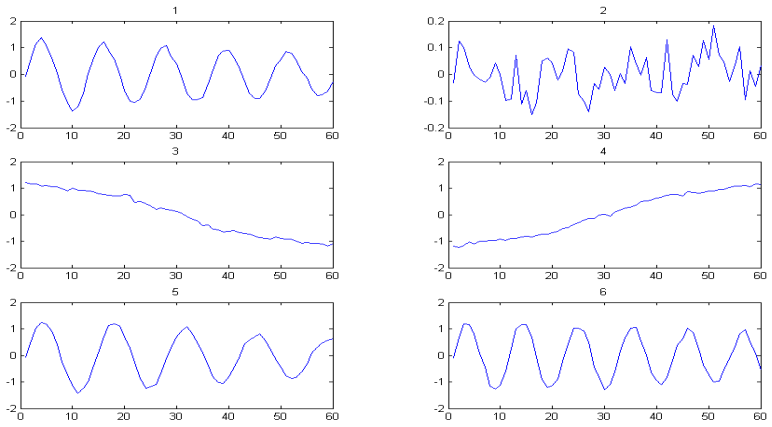
όπου c είναι ο αριθμός των επιπέδων ανάλυσης

ομοιότητα @ επίπεδο i

- **Μην έχοντας προηγούμενη γνώση ίσο βάρος σε όλα τα επίπεδα ανάλυσης αποδίδει καλά στις περισσότερες περιπτώσεις**

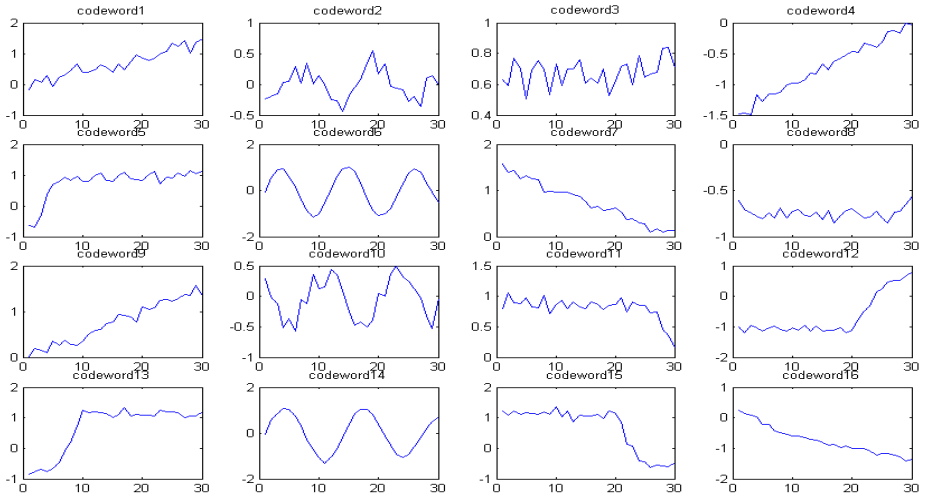
MVQ: Παράδειγμα Βιβλίου Κώδικα (Codebook)

- Πρώτο επίπεδο κωδικοποίησης



- Δεύτερο επίπεδο κωδικοποίησης

(περισσότεροι κωδικοί μιας και κωδικοποιείται περισσότερη λεπτομέρεια)

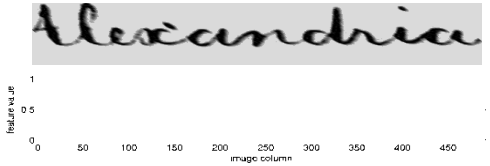


Πειράματα

Datasets:

- SYNDATA (control chart data): συνθετικό σετ :
  - 600 παραδείγματα διαγραμμάτων μήκους 60 σημείων που ανήκουν σε 6 διαφορετικές κατηγορίες (Normal, Cyclic, Increasing trend, Decreasing trend, Upward shift, and Downward shift)

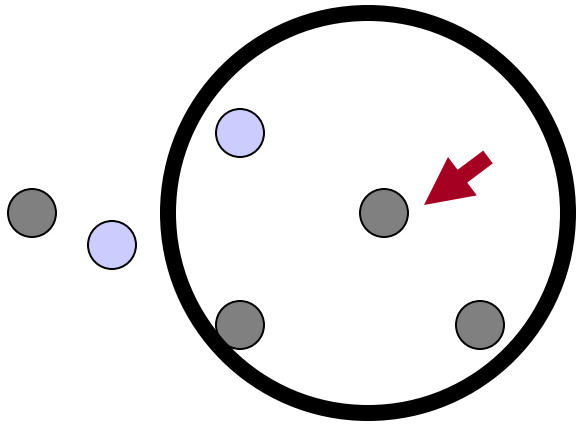
- CAMMOUSE: ένα χωροχρονικό σετ δεδομένων
  - Αποτελείται από 5 λέξεις και έχει δημιουργηθεί χρησιμοποιώντας το Camera Mouse Program, 3¨εγγραφές κάθε λέξης, μέσος όρος μήκους: 1100 σημεία



- RTT: RTT μετρήσεις από UCR στο CMU με ρυθμό αποστολής 50 msec για μια μέρα
  - Ο συνολικός αριθμός των τιμών RTT είναι 1,728,000. Το σετ δεδομένων διαμερίστηκε σε 24 χρονοσειρές μήκους 72,000.

Πειράματα

Αναζήτηση Best Match:



Matching accuracy: ποσοστό των knn's (που βρέθηκαν με διαφορετικές μεθόδους) που είναι στην ίδια κλάση

$$\text{Accuracy} = \frac{|\,\text{knn}(q) \cap \text{std\_set}(q)|}{k} \times 100\%$$

Πειράματα

Αναζήτηση Best Match

SYNDATA

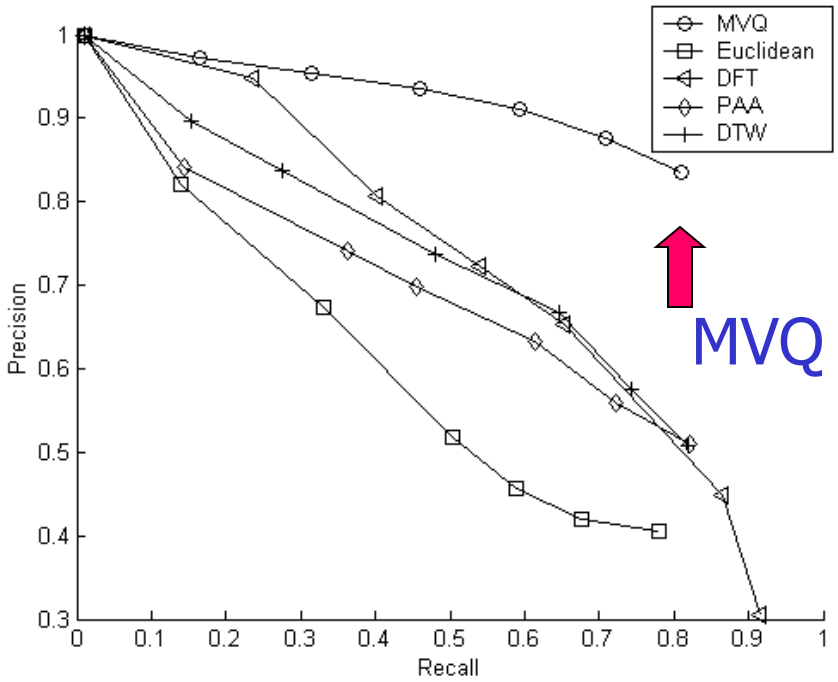
Method	Weight Vector	Accuracy
Single level VQ	[1 0 0 0 0]	0.55
	[0 1 0 0 0]	0.70
	[0 0 1 0 0]	0.65
	[0 0 0 1 0]	0.48
	[0 0 0 0 1]	0.46
MVQ	[1 1 1 1 1]	**0.83**
Euclidean		0.51

CAMMOUSE

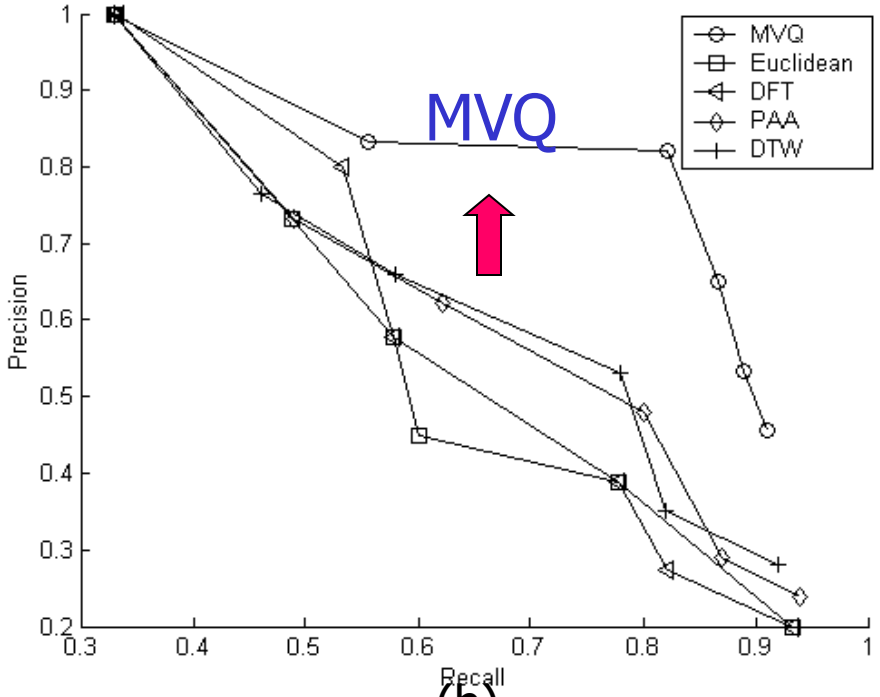
Method	Weight Vector	Accuracy
Single level VQ	[1 0 0 0 0]	0.56
	[0 1 0 0 0]	0.60
	[0 0 1 0 0]	0.44
	[0 0 0 1 0]	0.56
	[0 0 0 0 1]	0.60
MVQ	[1 1 1 1 1]	**0.83**
Euclidean		0.58

Πειράματα

Αναζήτηση Best Match



(a)



(b)

Precision-recall για διαφορετικές μεθόδους
(a) SYNDATA dataset (b) CAMMOUSE dataset

Πειράματα

Πειράματα ομαδοποίησης (clustering)

Δεδομένων δύο ομαδοποιήσεων (clusterings), $G=G_1, G_2, \ldots, G_K$ (οι αληθινοί clusters), και $A = A_1, A_2, \ldots, A_k$ (αποτέλεσμα ομαδοποίησης μιας μεθόδου), η ακρίβεια ομαδοποίησης (clustering accuracy) υπολογίζεται με την cluster similarity που ορίζεται ως εξής:

$$\mathrm{Sim}(G, A) = \frac{\sum_i \max_j Sim(G_i, A_j)}{k}$$

όπου

$$\mathrm{Sim}(Gi, Aj) = \frac{2|G_i \cap A_j|}{|G_i| + |A_j|}$$

[Gavrilov, M., Anguelov, D., Indyk, P. and Motwani, R., KDD 2000]

Πειράματα

Πειράματα ομαδοποίησης (clustering)

SYNDATA

Method	Weight Vector	Accuracy
Single level VQ	[1 0 0 0 0]	0.69
	[0 1 0 0 0]	0.71
	[0 0 1 0 0]	0.63
	[0 0 0 1 0]	0.51
	[0 0 0 0 1]	0.49
MVQ	[1 1 1 1 1]	**0.82**
DFT		0.67
SAX		0.65
DTW		0.80
Euclidean		0.55

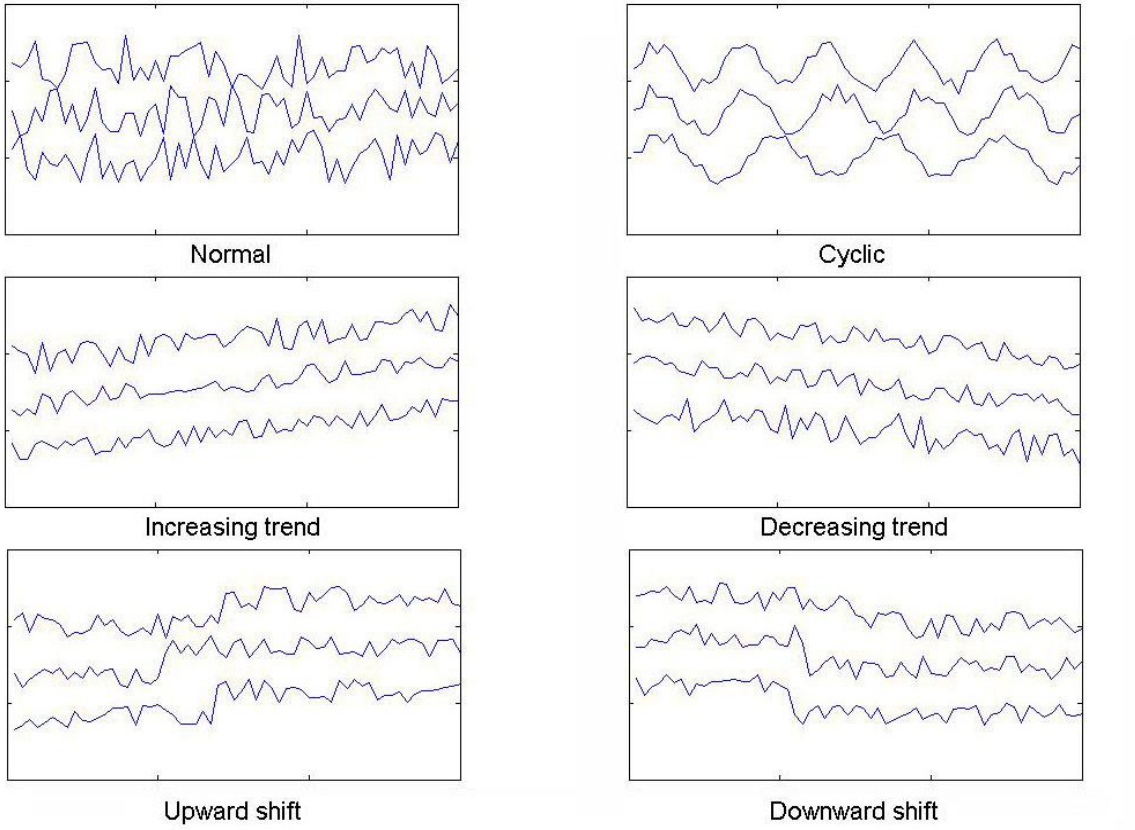
RTT

Method	Weight Vector	Accuracy
Single level VQ	[1 0 0 0 0]	0.55
	[0 1 0 0 0]	0.52
	[0 0 1 0 0]	0.57
	[0 0 0 1 0]	0.80
	[0 0 0 0 1]	0.79
MVQ	[0 0 0 1 1]	**0.81**
DFT		0.54
SAX		0.54
DTW		0.62
Euclidean		0.50

Πειράματα

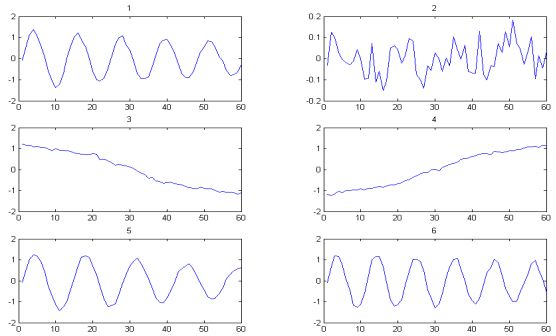
Δημιουργία περίληψης - Summarization (SYNDATA)

Τυπικές σειρές:



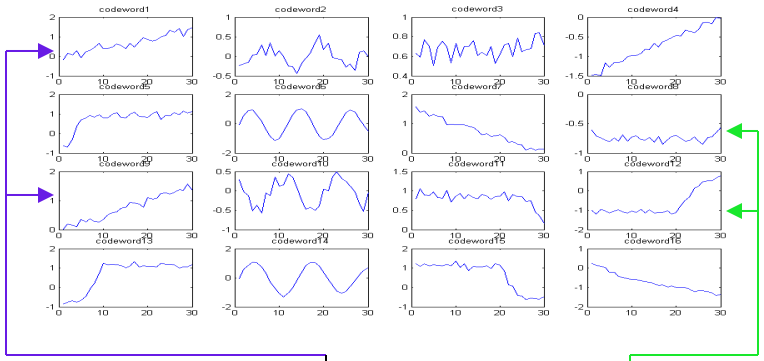
Πειράματα

Πρώτο επίπεδο



c	class1	class2	class3	class4	class5	class6
1	0, 0	100,31	0, 0	0, 0	0, 0	0, 0
2	96,100	4, 4	0, 0	0, 0	0, 0	0, 0
3	0, 0	0, 0	0, 0	50,100	0, 0	50,100
4	0, 0	0, 0	50,100	0, 0	50,100	0, 0
5	0, 0	100,40	0, 0	0, 0	0, 0	0, 0
6	0, 0	100,25	0, 0	0, 0	0, 0	0, 0

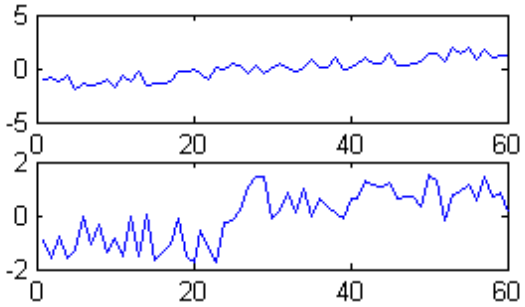
Δεύτερο επίπεδο



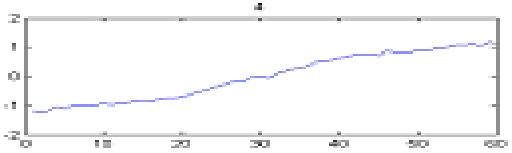
c	class1	class2	class3	class4	class5	class6
1	2, 1	0, 0	94, 19	0, 0	2, 1	2, 1
2	90, 51	10, 6	0, 0	0, 0	0, 0	0, 0
3	1, 1	0, 0	0, 0	3, 1	56, 21	40, 15
4	0, 0	0, 0	96, 48	0, 0	3, 1	1, 1
5	0, 0	0, 0	37, 5	0, 0	63, 9	0, 0
6	11, 5	89, 39	0, 0	0, 0	0, 0	0, 0
7	0, 0	0, 0	0, 0	100, 45	0, 0	0, 0
8	0, 0	0, 0	1, 1	3, 2	46, 28	5, 29
9	0, 0	0, 0	98, 26	0, 0	2, 1	0, 0
10	78, 39	22, 11	0, 0	0, 0	0, 0	0, 0
11	0, 0	0, 0	0, 0	12, 3	25, 7	63, 19
12	0, 0	0, 0	7, 1	0, 0	93, 21	0, 0
13	0, 0	0, 0	0, 0	0, 0	100, 11	0, 0
14	4, 2	96, 44	0, 0	0, 0	0, 0	0, 0
15	0, 0	0, 0	0, 0	3, 1	0, 0	97, 15
16	1, 1	0, 0	0, 0	70, 48	0, 0	29, 20

MVQ: Παράδειγμα: Δυο Χρονικές Σειρές

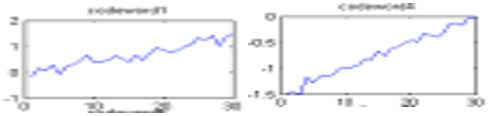
- Δεδομένων δυο χρονικών σειρών t1 και t2 ως εξής:

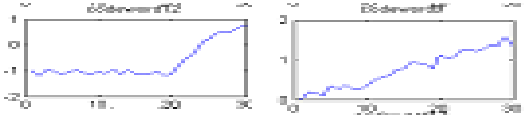


- Στο πρώτο επίπεδο, κωδικοποιούνται με τον ίδιο κωδικό (codeword) (3), και δεν υπάρχει διαφορά ανάμεσα στις δυο σειρές



- Στο δεύτερο επίπεδο, καταγράφονται περισσότερες λεπτομέρειες. Οι δυο σειρές έχουν διαφορετική κωδικοποίηση: η πρώτη κωδικοποιείται με τους κωδικούς 1 και 4, η δεύτερη με τους κωδικούς 9 and 12.





Πολυπλοκότητα προεπεξεργασίας

- Time complexity:
  - T(training) = $c * i * s * N * w = O(N * w)$
    - N: αριθμός χρονικών σειρών στο training set
    - w: αριθμός τμημάτων στη μεγαλύτερη ανάλυση
    - N*w: αριθμός training vectors
    - c: αριθμός των επιπέδων ανάλυσης (~5)
    - i: αριθμός επαναλήψεων Lloyd (~15)
    - s: μέγεθος του βιβλίου κώδικα (~ 6-32)
- Η μάθηση του βιβλίου κώδικα γίνεται μια φορά κατά τη διάρκεια της προεπεξεργασίας

Συμπεράσματα

- Μια νέα αναπαράσταση χρονικών σειρών που χρησιμοποιεί σύμβολα/κώδικες και ανάλυση πολλαπλών επιπέδων
- Νέα συνάρτηση απόστασης
- Βελτιωμένη αποδοτικότητα βασισμένη σε μείωση της διαστατικότητας
- Εργαλεία για δημιουργία περίληψης
- Εκμετάλλευση υπάρχουσας γνώσης (prior knowledge) σχετικά με τα δεδομένα

Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Έλεγχος/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Ανάλυση Ομοιότητας Χρονικών Σειρών Διαφορετικού Μήκους

- Πρόβλημα - Κίνητρα
- MVM (Minimal Variance Matching)
  - περιγραφή, παράδειγμα
- Πειραματικά αποτελέσματα
  - Ταίριασμα ολόκληρων σειρών (whole sequence matching)
  - Ταίριασμα υπο-σειρών (subsequence matching)
- Συμπεράσματα

Πρόβλημα – Κίνητρα

- Πρόβλημα: Δεδομένων 2 χρονικών σειρών:

  $A = \{ a_1, \ldots a_m \}$ (query) και

  $B = \{ b_1, \ldots, b_n \}$ (target),

προσδιορίστε υπο-σειρά B' της B που ταιριάζει καλύτερα με την A (όπου $m < n$)

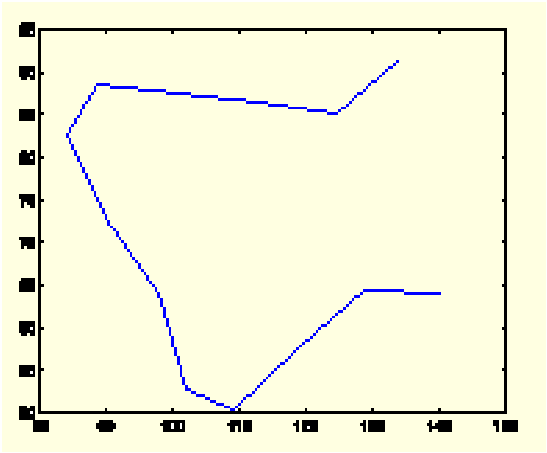
- Για κάποια δεδομένα μπορούμε εύκολα και με ακρίβεια να εξάγουμε την αρχή και το τέλος κάποιου προτύπου (pattern) εντός μίας σειράς (π.χ., καρδιακοί παλμοί). Σε μερικές εφαρμογές όμως αυτό είναι δύσκολο.

Κίνητρα (συνέχεια)

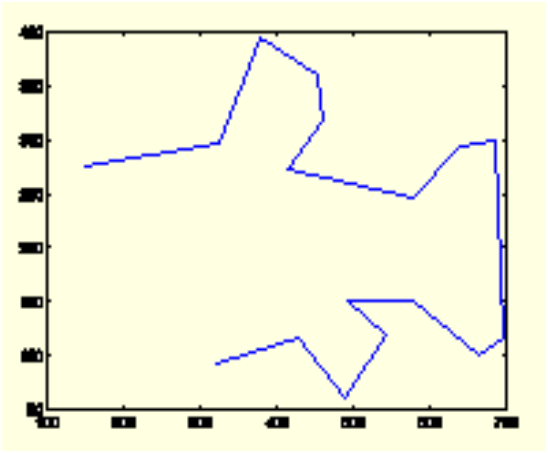
Παράδειγμα:

Δεδομένου ενός τμήματος κάποιου σχήματος, θέλουμε να βρούμε παρεμφερή σχήματα που το περιέχουν

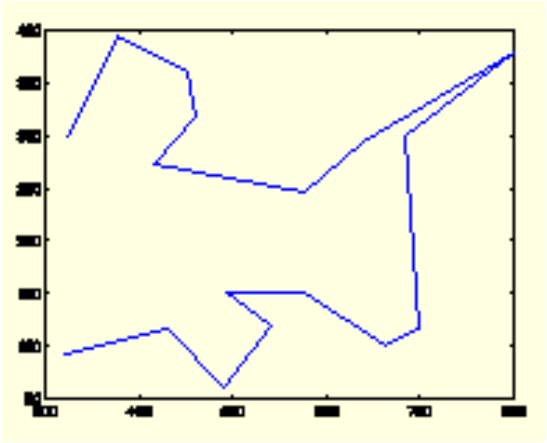
Query Shape



Target Shape

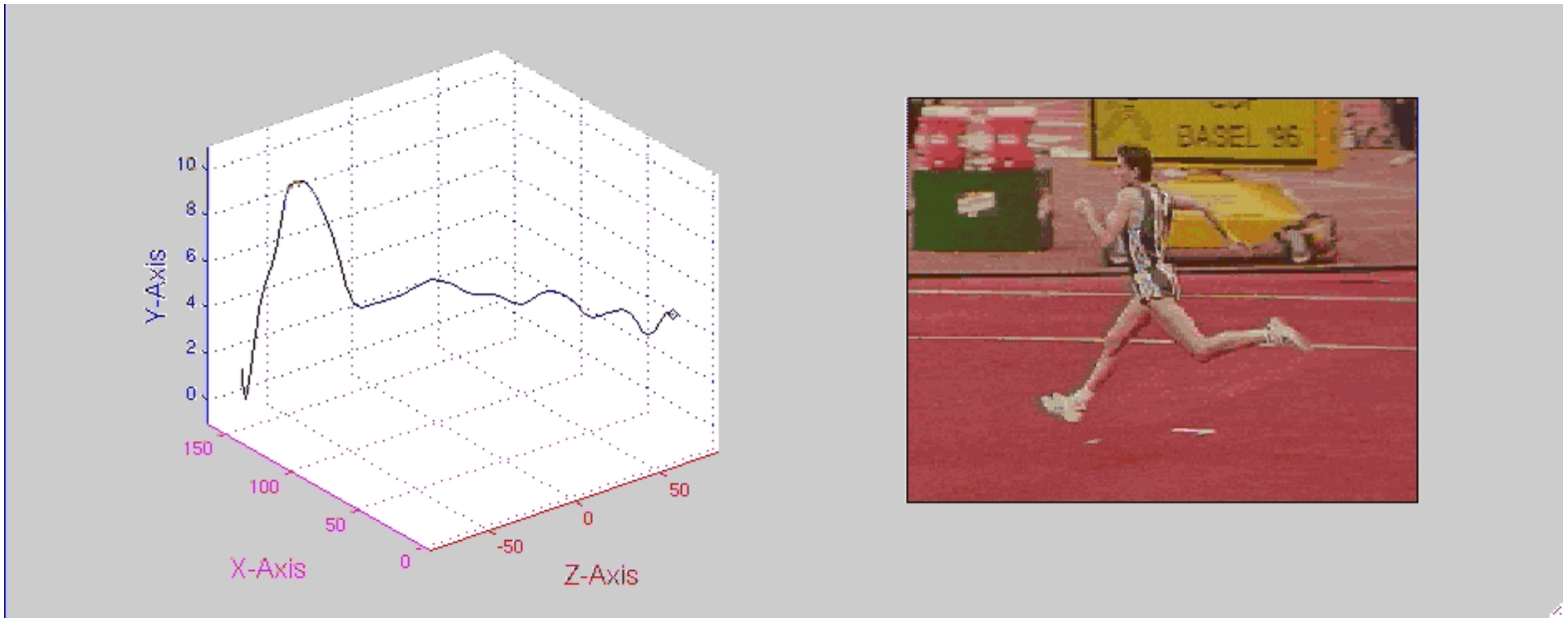


Target Shape



Κίνητρα (συνέχεια)

Ένα άλλο παράδειγμα:



Αναπήδηση αθλητή άλματος εις ύψος

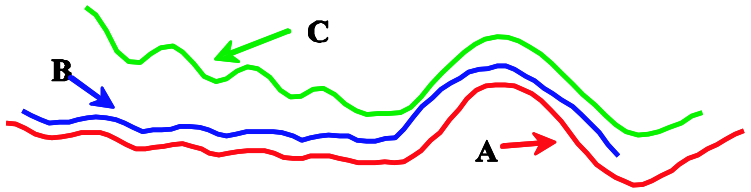
Άλλες μέθοδοι

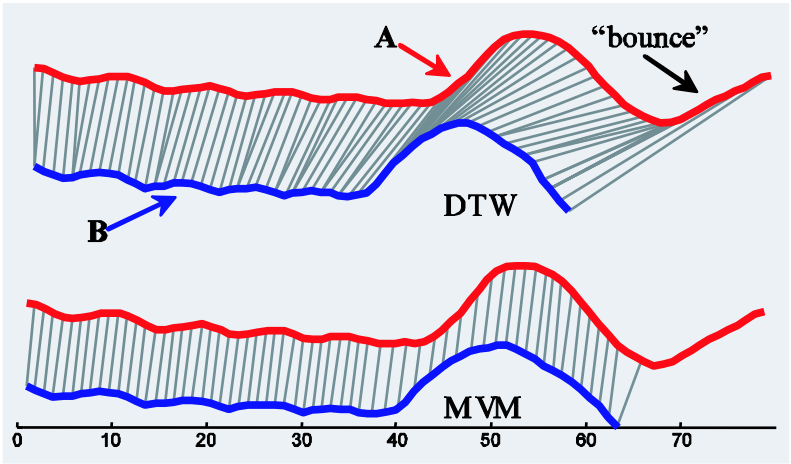
- Dynamic Time Warping (DTW):
  - Ευθυγραμμίζει τον άξονα χρόνου πριν από τον υπολογισμό της απόστασης των δύο χρονικών σειρών
  - Η απόσταση είναι το άθροισμα των αποστάσεων των αντίστοιχων στοιχείων
  - Χρησιμοποιεί dynamic programming για να βρεί την αντιστοιχία έτσι ωστε να ελαχιστοποιηθεί η απόσταση
  - Ευαίσθητη στα outliers μιας και δεν μπορεί να παραλείψει κανένα στοιχείο της target σειράς
- Longest Common Subsequence (LCSS):
  - Προσδιορίζει την μεγαλύτερη κοινή υπο-σειρά
  - Η απόσταση είναι το ratio (μήκος της μεγαλύτερης κοινής υπο-σειράς) / (μήκος της σειράς)
  - Όχι απαραίτητα συνεχόμενα σημεία, η σειρά των στοιχείων δεν αλλάζει και κάποια στοιχεία μπορούν να παραλειφθούν (no outlier problem)
  - Χρειάζεται ένα κατώφλι, θ, που καθορίζει πότε οι αριθμητικές τιμές είναι ίσες

Κίνητρα (συνέχεια)

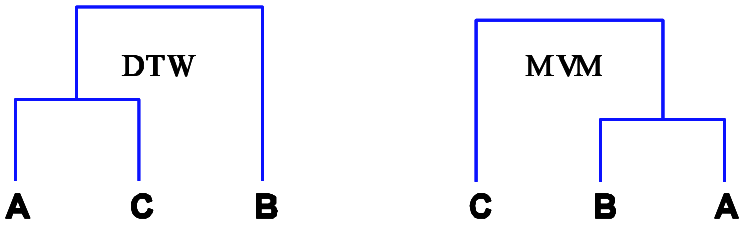
A, B: ένα ψηλός αθλητής
(η αναπήδηση στο B περικόπτεται)

C: σχετικά κοντή αθλήτρια με αρκετά διαφορετικό style





- Το DTW αντιστοιχεί την αναπήδηση του A στο τέλος της σειράς B
- Θέλουμε να μπορούμε να αγνοούμε τμήματα που δεν έχουν φυσική αντιστοιχία
- Η προτεινόμενη μέθοδος, MVM, παράγει πιο φυσικές ομάδες (clustering) αφού επιτρέπει την εξάλειψη των "outliers"



Minimal Variance Matching (MVM)

Περιγραφή της μεθόδου

- MVM: αλγόριθμος για εύρεση ελαστικού ταιριάσματος (elastic matching) δύο χρονικών σειρών A,B διαφορετικών μηκών m, n
- Στόχος είναι να βρούμε την καλύτερη αντιστοιχία (correspondence) της σειράς A σε μια υπο-σειρά B' της B
- Η *αντιστοιχία* ορίζεται ως εξής:
  - a monotonic injection $f : \{1,\ldots,m\} \rightarrow \{1,\ldots,n\}$ (όπου $f(i) < f(i+1)$)

  έτσι ώστε a_i αντιστοιχίζεται στο $b_{f(i)}$ για κάθε i στο {1,…,m}
  - δεν είναι απαραίτητο να συμμετέχουν όλα τα στοιχεία του B
- Το σύνολο των δεικτών $\{ f(1), \ldots , f(m) \}$ ορίζει την υπο-σειρά B' της B
- Στο DTW η αντιστοιχία είναι μια σχέση one-to-many και many-to-one στο σύνολο των δεικτών {1,…,m} x {1,…,n}.
- Από όλες τις πιθανές αντιστοιχίες ψάχνουμε την “καλύτερη”

[L. J. Latecki, V. Megalooikonomou, Q. Wang, R. Lakaemper, C. A. Ratanamahatana, and E. Keogh, "Elastic Partial Matching of Time Series", *PKDD'05*]

Minimal Variance Matching (MVM)

Περιγραφή της μεθόδου

- Η "καλύτερη" αντιστοιχία ορίζεται ως εξής (για A,B με μήκος m,n όπου m<n):
  - Είναι η αντιστοιχία *f** που δίνει το global minimum της Ευκλείδειος απόστασης D μεταξύ $(a_i, b_{f^*(i)})$ για όλα τα i στο {1,…,m}
  - $$d(a,b,f) = \sqrt{\sum_{i=1}^{m} (b_{f(i)} - a_i)^2}$$
  - Το σύνολο των δεικτών { *f**(1), … , *f**(m) } ορίζει το "καλύτερο" ταίριασμα της υπο-σειράς B' της B
- Από την πλευρά της στατιστικής:
  η σειρά A είναι μια έκδοση της B' με κάποιο θόρυβο: $A \sim B' + N(0,\sigma^2)$
  ($N(0, \sigma^2)$ είναι μια zero-mean Gaussian noise μεταβλητή)

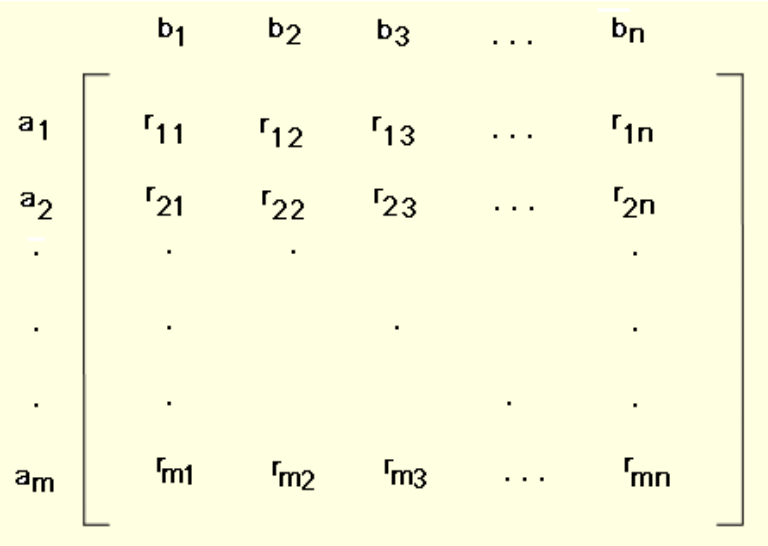
$$\sigma^2(a,b,f) = \frac{1}{m}\sum_{i=1}^{m} (b_{f(i)} - a_i)^2$$

Minimal Variance Matching (MVM)

Περιγραφή της μεθόδου

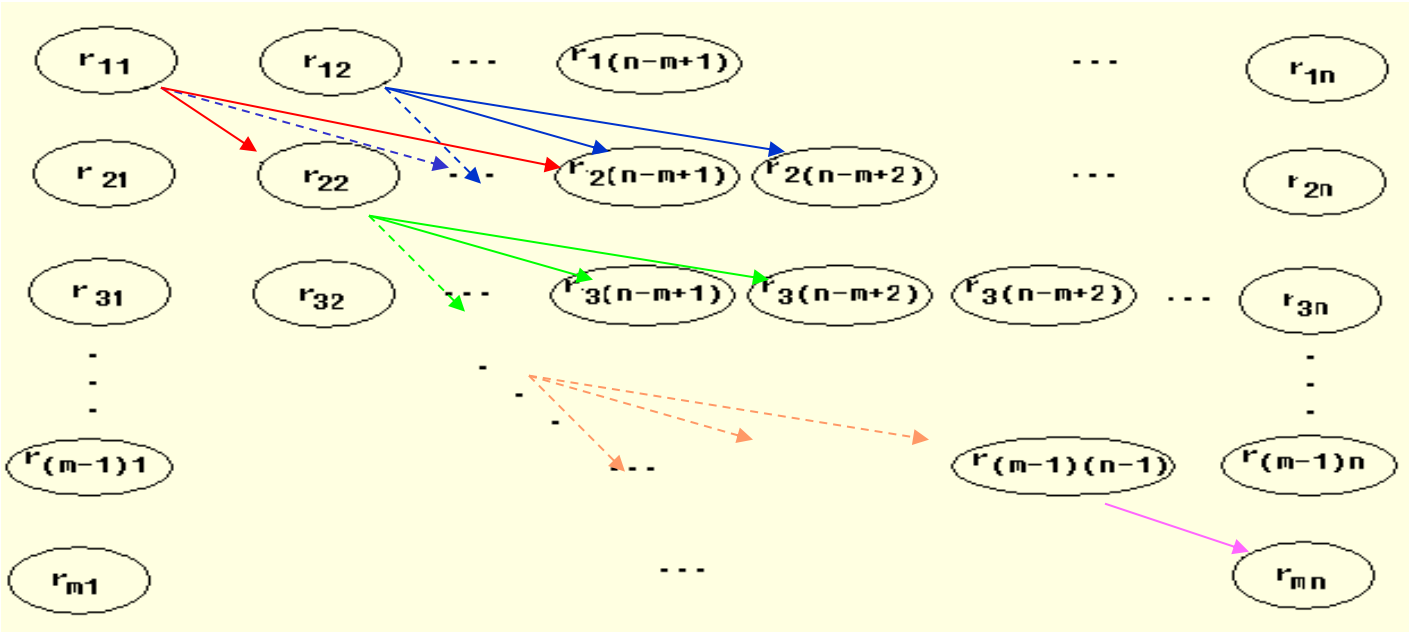
Από τον πίνακα αποστάσεων (**R**) μεταξύ **A** (a_1, a_2, …,a_m) και **B** (b_1, b_2, …, b_n)

$\mathbf{R} = \{\mathbf{r_{ij}} \mid r_{ij} = a_i - b_j;\ 1 \le i \le m \ \&\ 1 \le j \le n\ \}$



Κατασκευάζουμε έναν Directed Acyclic Graph (DAG) →

Minimal Variance Matching (MVM)



- Οι κορυφές (vertices) είναι στοιχεία του R (πίνακα αποστάσεων)
- Οι γραμμές (edges) υπάρχουν μεταξύ r_{ij} $(1 \le i \le m;\ i \le j)$ και r_{kp} $(k \le p)$ **αν και μόνο αν**

  $k-i = 1$ (γραμμή προς γραμμή) και

  $j+1 \le p \le j+1 + (n - m) - (j - i)$ (μπορούμε να παραλείψουμε στήλες)
- Συνάρτηση κόστους (cost function): edge weight μεταξύ r_{ij} και r_{kp} : $weight(r_{ij},r_{kp}) = (r_{kp})^2$
- Ψάχνουμε για το shortest path από r_{1x} $(1 \le x \le n-m+1)$ στο r_{mv} $(m \le v \le n)$

Minimal Variance Matching (MVM)

Ένα απλό παράδειγμα

- A = (1, 2, 8, 6, 8) B = (1, 2, 9, 3, 3, 5, 9)
- Για να υπολογίσουμε την optimal αντιστοιχία *f**, αναπτύσσουμε τον παρακάτω πίνακα διαφορών:

$$\begin{bmatrix} 0 & 1 & 8 & 2 & 2 & 4 & 8 \\ -1 & 0 & 7 & 1 & 1 & 3 & 7 \\ -7 & -6 & 1 & -5 & -5 & -3 & 1 \\ -5 & -4 & 3 & -3 & -3 & -1 & 3 \\ -7 & -6 & 1 & -5 & -5 & -3 & 1 \end{bmatrix}$$

Minimal Variance Matching (MVM)

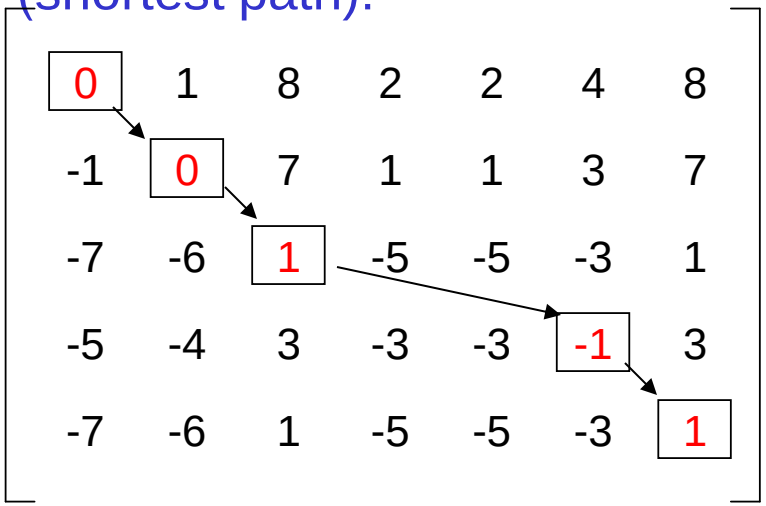
Ένα απλό παράδειγμα

- A = (1, 2, 8, 6, 8) B = (1, 2, 9, 3, 3, 5, 9)
- Το συντομότερο μονοπάτι (shortest path):

$f^*(1) = 1$, $f^*(2) = 2$,

$f^*(3) = 3$, $f^*(4) = 6$,

$f^*(5) = 7$



- Η μικρότερη απόσταση μεταξύ A και B είναι sqrt(3) ≈ 1.732

Πειραματικά Αποτελέσματα

- Πειράματα:
  - Ταίριασμα ολόκληρων σειρών (whole sequence matching) - διάφορα benchmark datasets
  - Ταίριασμα υπο-σειρών (subsequence matching) (ομοιότητα σχημάτων)

  - Μέτρηση ακρίβειας ταξινόμησης (classification accuracy)
  - Σύγκριση με άλλες μεθόδους (LCSS & DTW)

[L. J. Latecki, V. Megalooikonomou, Q. Wang, R. Lakaemper, C. A. Ratanamahatana and E. Keogh, "Partial Elastic Matching of Time Series", *ICDM'05*]

Πειραματικά Αποτελέσματα (συνέχεια)

Ταίριασμα ολόκληρων σειρών (Whole sequence matching)

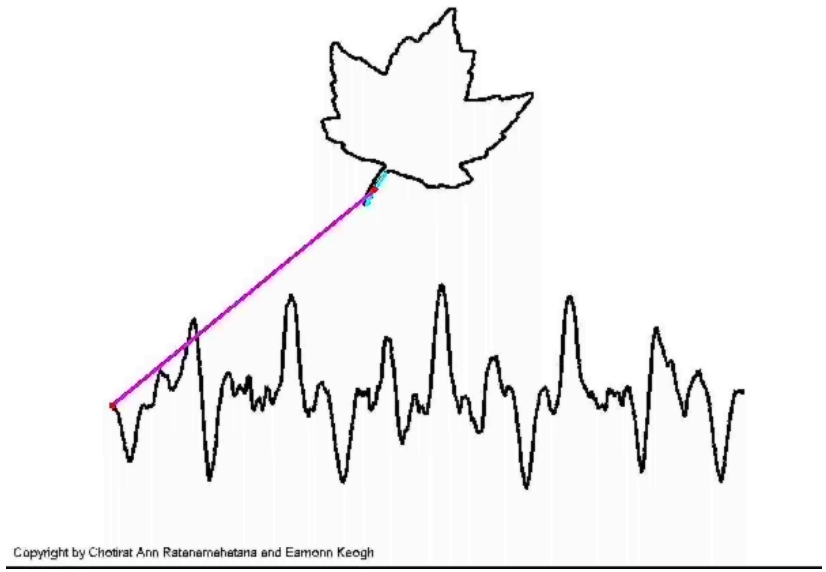
3 σετ δεδομένων:

Face: 112 σειρές που απεικονίζουν το προφίλ 4 διαφορετικών ανθρώπων
- 107 – 240 σημεία

Leaf: 442 σειρές που απεικονίζουν το περίγραμμα 6 διαφορετικών ειδών φύλλων
- 22 – 475 σημεία

Gun: 200 κινήσεις τραβήγματος πιστολιού από 2 διαφορετικούς ηθοποιούς
- 150 σημεία

Animation της μετατροπής
"σχήματος σε χρονική σειρά"



Πειραματικά Αποτελέσματα (συνέχεια)

Ταίριασμα ολόκληρων σειρών – Αποτελέσματα

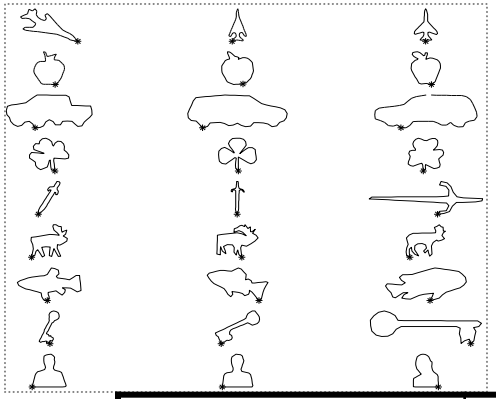
1-NN	Face	Gun	Leaf
MVM	98.21	100	97.29
DTW	96.43	99.00	96.38
LCSS	97.32	99.50	92.53

- Η καλύτερη απόδοση του MVM βασίζεται στην ικανότητα του να αποκλείει αυτόματα στοιχεία της target σειράς που δεν ταιριάζουν (π.χ., πίσω μέρος του κεφαλιού στο προφίλ)

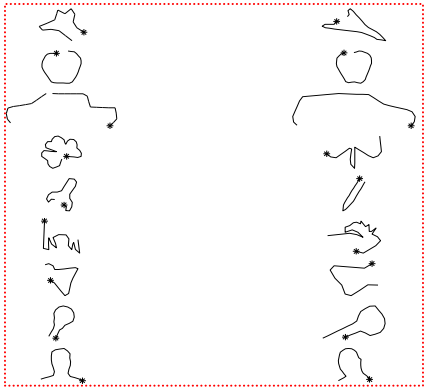
Πειραματικά Αποτελέσματα (συνέχεια)

Ταίριασμα υπο-σειρών (ομοιότητα σχημάτων)

Σχήματα



Μέρη σχημάτων που χρησιμοποιούνται σαν queries



	1-NN	in best 6
MVM	85	60
LCSS	75	43
DTW	15	30

Το MVM προσδιορίζει την καλύτερη αντιστοιχία υποσειράς στο σχήμα

Πολυπλοκότητα

- Time complexity:
  - MVM
    - O(m * (n-m) * (n-m)) = O(m*n^2) [το (n-m) είναι λόγω του περιορισμού στην ολική ελαστικότητα]
    - O(m * (n-m) * c) = O(m*n) [Όταν περιορίσουμε τον δείκτη διαφοράς σε σταθερά *c*]
    - O(m) [οταν περιορίσουμε τη διαφορά μήκους των δύο σειρών σε μια σταθερά K: n-m ≤ K
  - DTW
    - O(m * n)
    - O(n) [με περιορισμό στο warping window]
  - LCSS
    - O(m * n)
- Space complexity:
  - MVM, DTW: Two matrices: one for cost, one for path, both are m * n.
  - LCSS: One matrix: record the length of LCS, m *n

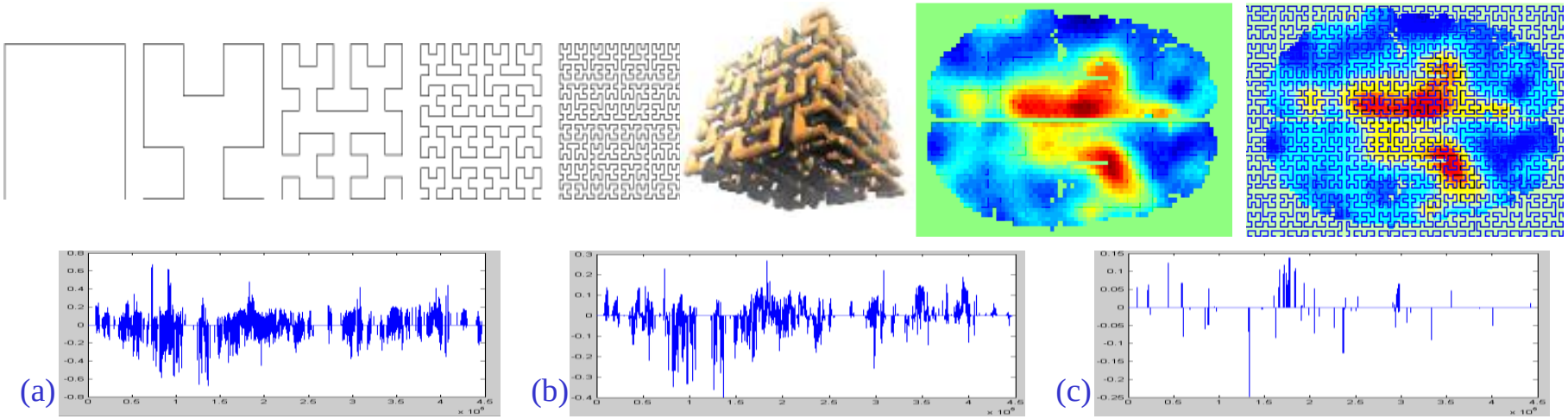
Συμπεράσματα

- Το MVM (Minimal Variance Matching) κάνει τα παρακάτω ταυτόχρονα :
  - Αυτόματα προσδιορίζει αν η query σειρά ταιριάζει καλύτερα με όλη την target σειρά, ή μόνο με μέρος της
  - Εντοπίζει την υπο-σειρά που ταιριάζει καλύτερα με την query σειρά
  - Αυτόματα παραλείπει τα outliers της target σειράς
  - Ελαχιστοποιεί την στατιστική διακύμανση της ανομοιότητας των αντίστοιχων στοιχείων
- Το MVM υπερτερεί των DTW και LCSS σε απόδοση
- Μετασχηματίζοντας το πρόβλημα του ελαστικού ταιριάσματος χρονικών σειρών στο πρόβλημα εύρεσης του πιο σύντομου μονοπατιού σε έναν γράφο DAG, είναι δυνατή η χρήση πιο αποδοτικών αλγορίθμων.

Κύρια σημεία

- Εισαγωγή
- Χωρικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning) και ομαδοποίηση (clustering)
  - Ανακάλυψη συσχετίσεων
  - Ανακάλυψη διακριτικών περιοχών
  - Εκτίμηση/αξιολόγηση της διαδικασίας εξόρυξης γνώσης
- Χρονικά δεδομένα
  - Κίνητρα – Προβλήματα – Υπόβαθρο
  - Διαμερισμός (partitioning)
  - Προσέγγιση σειρών με Multi-resolution Vector Quantization
  - Ανάλυση ομοιότητας χρονικών σειρών διαφορετικού μήκους
- Εφαρμογή τεχνικών ανάλυσης σειρών σε χωρικά δεδομένα
- Συμπεράσματα – Συζήτηση

Ανάλυση εικόνων με μετασχηματισμό σε 1D

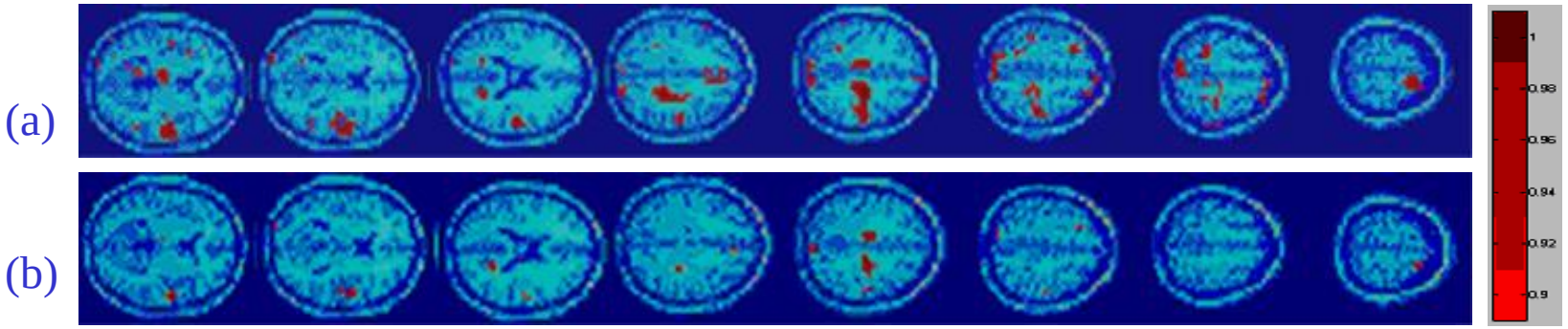


(a) linear mapping of a 3D fMRI scan, (b) effect of binning by representing each bin with its V_{mean} measurement, (c) the discriminative voxels after applying the t-test with θ=0.05

- Hilbert Space Filling Curve
- Τοποθέτηση σε κουτιά (Binning)
- Statistical tests of significance σε **ομάδες από σημεία**
- Προσδιορισμός περιοχών ικανών να διακρίνουν ομάδες (με back-projection)

[D. Kontos, V. Megalooikonomou, N. Ghubade, and C. Faloutsos, "Detecting discriminative functional MRI activation patterns using space filling curves", IEEE Engineering in Medicine and Biology Society (EMBS), 2003]

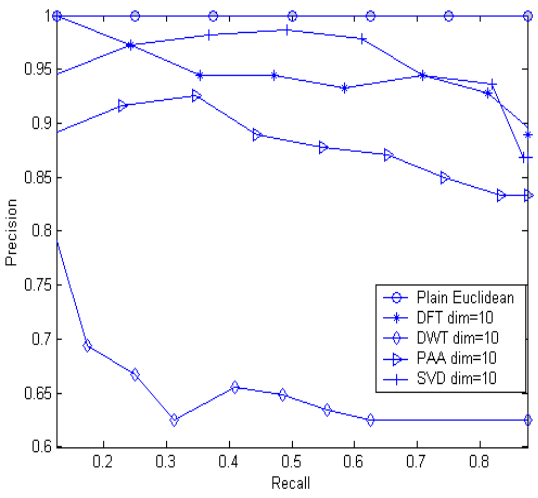
Χρήση τεχνικών για χρονικές σειρές



Areas discovered: (a) θ=0.05, (b) θ=0.01. The colorbar shows significance.

Αποτελέσματα: 87%-98% ακρίβεια ταξινόμησης (t-test, CATX)

<u>Παραλλαγή:</u> Αλυσιδωτή σύνδεση των τιμών των σημαντικών περιοχών → χωρικές σειρές



- Ανάλυση προτύπων με βάση την ομοιότητα μεταξύ χωρικών σειρών και χρονικών σειρών
  - SVD, DFT, DWT, PCA (ακρίβεια ομαδοποιήσης - clustering accuracy: 89-100%)

[Q. Wang, D. Kontos, G. Li and V. Megalooikonomou, " Application of Time Series Techniques to Data Mining and Analysis of of Spatial Patterns in 3D images", ICASSP 2004]

Περίληψη και Γενικά Συμπεράσματα

- Στόχος: ‘Ανεύρεση προτύπων (patterns) / ενδιαφέροντων πραγμάτων’ αποτελεσματικά και με robustness σε Βάσεις χωρικών και χρονικών δεδομένων
- Χρήση διαμερισμού και ομαδοποίησης
- Χρήση πολλαπλής ανάλυσης (multiple resolutions)
- Ελαχιστοποίηση του αριθμού των tests που εκτελούνται
- Επιλεκτική επεξεργασία του χώρου με έξυπνες μεθόδους για ανεύρεση διακριτικών περιοχών (discriminative areas)
- Μείωση της διαστατικότητας
- Αναπαράσταση με συμβολοσειρές
- Εργαλεία για δημιουργία περίληψης
- Εκμετάλλευση υπάρχουσας γνώσης (prior knowledge) σχετικά με τα δεδομένα
- Μελέτη πιο πολύπλοκων χωρικών/χρονικών/χωροχρονικών μοντέλων και μεγαλύτερων σετ δεδομένων