

Μάθηση από δεδομένα πολλαπλών ετικετών Multi-label Learning

Κάτια Κερμανίδου
kerman@ionio.gr

Multi-label Learning

- Το πρόβλημα διέπεται από περισσότερες της μίας μεταβλητές-στόχους (εξόδου)
- Οι μεταβλητές εξόδου είναι δυαδικές
- Παράδειγμα: επισημείωση φωτογραφίας



*Partylife, Citylife, Outdoor,
Landscape_Nature, Clouds, Sky,
Day, No_Visual_Season, Plants,
Trees, Neutral_Illumination,
No_Blur, Big_Group, Cute,
Female, Male, Adult*

Συμβολισμοί

Έστω d το πλήθος των ανεξάρτητων μεταβλητών, X

- Οι ανεξάρτητες μεταβλητές μπορούν να είναι ονοματικές ή αριθμητικές

Έστω q το πλήθος των ετικετών εξόδου $\mathcal{Y} = \{\lambda_1, \lambda_2, \dots, \lambda_q\}$

Έστω ένα σετ δεδομένων πολλαπλών ετικετών αποτελούμενο από n παραδείγματα εκπαίδευσης

Έξοδος:

A. Ταξινόμηση: η διαίρεση του σετ ετικετών σε

- ένα σχετικό (θετικό) υπο-σετ, και
- ένα μη-σχετικό (αρνητικό) υπο-σετ για ένα παράδειγμα ελέγχου.

B. Κατάταξη: η κατάταξη όλων των ετικετών ανάλογα με την σχετικότητά τους για ένα παράδειγμα ελέγχου

Γ. Συνδυασμός των A και B

Δ. Υπολογισμός πιθανότητας σχετικότητας κάθε ετικέτας για ένα παράδειγμα ελέγχου

$$p(\lambda_1|x) = 0.8, p(\lambda_2|x) = 0.2$$

$$p(\lambda_3|x) = 0.9, p(\lambda_4|x) = 0.4, p(\lambda_5|x) = 0.1$$

Παραδείγματα εφαρμογών

- Ταξινόμηση κειμένων
 - Ένα άρθρο πάνω στον μηχανισμό των Αντικυθήρων μπορεί να ταξινομηθεί σαν επιστήμη/τεχνολογία/ιστορία/πολιτισμός
 - Ένα επιστημονικό άρθρο πάνω στις συλλογικές μεθόδους μάθησης για εξόρυξη σε βιολογικά δεδομένα μπορεί να ταξινομηθεί σαν εξόρυξη βιολογικών δεδομένων / συλλογική μάθηση
- Ταξινόμηση ήχου
 - Ένα κομμάτι ταξινομείται ως προς την διάθεση, το ύφος, τα βασικά μουσικά όργανα, τον ρυθμό, την γλώσσα κλπ
- Ταξινόμηση εικόνας
- Ταξινόμηση βιολογικών δεδομένων
 - Ιεραρχίες οικογενειών πρωτεϊνών
- Επισημειώσεις σε ψηφιακά αντικείμενα στο web 2.0 από χρήστες
- Ταξινόμηση φαρμάκων
- Ιατρική διάγνωση
- Συσχέτιση καταναλωτών με πολλαπλές κατηγορίες προϊόντων από εταιρίες μάρκετινγκ

Αξιολόγηση example-based

- Έστω ότι
 - m το σύνολο των παραδειγμάτων ελέγχου
 - P_i είναι το σετ των προβλεφθέντων ετικετών για το παράδειγμα x_i
 - Y_i είναι το σετ των πραγματικών ετικετών του x_i
- Exact Match Ratio
 $\frac{1}{m} \sum_{i=1}^m I(P_i = Y_i)$, where $I(\text{true})=1$ and $I(\text{false})=0$
 $= 1/5 * (1+0+0+1+0) = 2/5$
- Ορθότητα: το ποσοστό των σωστά προβλεφθέντων ετικετών προς το σύνολο των ετικετών (πραγματικών & προβλεφθέντων) για το παράδειγμα (μ. όρος για όλα τα παραδείγματα)
 $= 1/5 * (1/1 + \frac{1}{2} + 0/2 + 2/2 + 1/2) = 3/5$
- Ακρίβεια: το ποσοστό των σωστά προβλεφθέντων ετικετών προς το σύνολο των προβλεφθέντων τιμών του παραδείγματος (μ. όρος για όλα τα παρ/τα)
 $= 1/5 * (1/1 + \frac{1}{2} + 0/1 + 2/2 + 1/1) = 3.5/5$
- Ανάκληση: το ποσοστό των σωστά προβλεφθέντων ετικετών προς το σύνολο των πραγματικών τιμών του παραδείγματος (μ. όρος για όλα τα παρ/τα)
 $= 1/5 * (1/1 + 1/1 + 0/1 + 2/2 + 1/2) = 3.5/5$

	TRUE	PREDICTED
1	A	A
2	B	A, B
3	B	A
4	A, B	A, B
5	A, B	B

Αξιολόγηση label-based

Όλα τα γνωστά μέτρα αξιολόγησης μπορούν να εφαρμοστούν εδώ

A. για κάθε ετικέτα ξεχωριστά και μετά να υπολογιστεί ο μέσος όρος για όλες τις ετικέτες (macro-averaging)

B. Καθολικά για όλα τα παραδείγματα και όλες τις ετικέτες (micro-averaging)

Contingency Table for λ_j		Actual Value	
		POS	NEG
Learner Output	POS	TP_j	FP_j
	NEG	FN_j	TN_j

Έστω ο διπλανός πίνακας σύγχυσης για την ετικέτα λ_j .

A. Υπολογίζω ορθότητα για την ετικέτα αυτή:

$$\text{Ορθότητα} = (TP_j + TN_j) / (TP_j + FP_j + TN_j + FN_j)$$

Και για κάθε ετικέτα και παίρνω τον μέσο όρο για όλες τις ετικέτες (Macro-ορθότητα).

B. Υπολογίζω το σύνολο (άθροισμα) των TPs σε όλες τις ετικέτες (TotalTPs), το σύνολο (άθροισμα) των TNs σε όλες τις ετικέτες (TotalTNs), το σύνολο (άθροισμα) των FPs σε όλες τις ετικέτες (TotalFPs), το σύνολο (άθροισμα) των FNs σε όλες τις ετικέτες (TotalFNs).

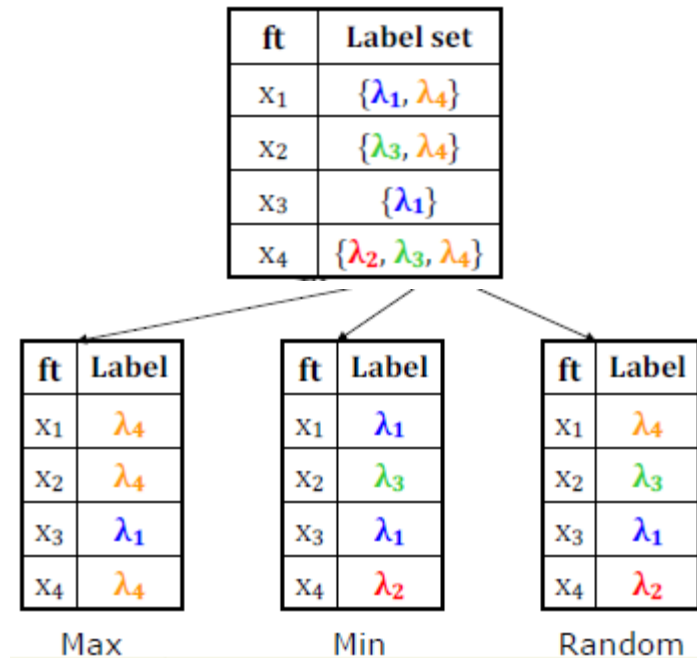
$$\text{MicroΟρθότητα} = (\text{TotalTPs} + \text{TotalTNs}) / (\text{TotalTPs} + \text{TotalTNs} + \text{TotalFPs} + \text{TotalFNs})$$

Μάθηση

- Α. Μετατρέπω το πρόβλημα σε ένα ή περισσότερα προβλήματα μάθησης μοναδικής ετικέτας
 - Αυτή η προσέγγιση είναι ανεξάρτητη αλγορίθμου μάθησης
- Β. Πειράζω τον αλγόριθμο μάθησης ώστε να μπορεί να αντιμετωπίσει προβλήματα πολλαπλών ετικετών ευθέως.

Μετατροπή του προβλήματος (1/5)

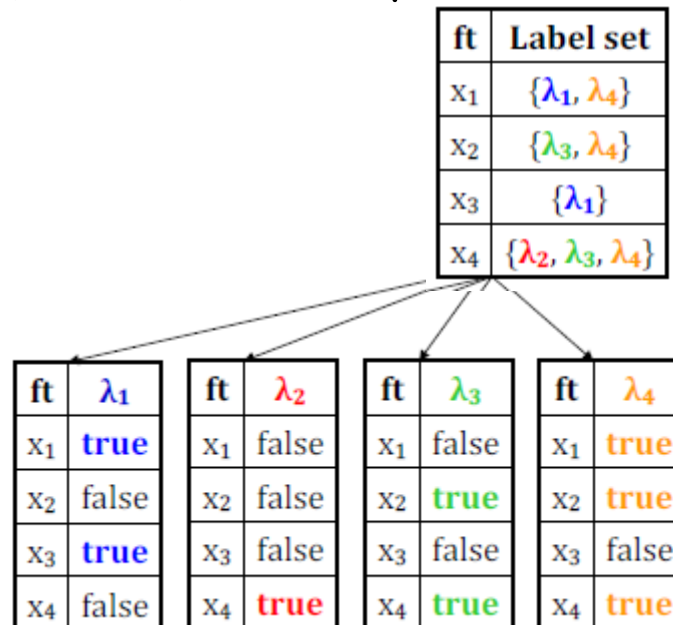
- Επιλογή μιας ετικέτας για κάθε παράδειγμα → **Απώλεια πληροφορίας**
 - Της πιο συχνής (Max)
 - Της πιο σπάνιας (Min)
 - Τυχαία επιλογή (random)
- Διαγραφή των παραδειγμάτων που έχουν πολλαπλές ετικέτες
 - Μεγάλη απώλεια πληροφορίας



Μετατροπή του προβλήματος (2/5)

- Έστω το σύνολο ετικετών $\mathcal{Y} = \{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$
- Binary Relevance
 - Κάθε ετικέτα θεωρείται μια δυαδική μεταβλητή
 - Πραγματοποιείται μάθηση κάθε ετικέτας ξεχωριστά (ξεχωριστός ταξινομητής)
 - Η έξοδος είναι η ένωση των διαφορετικών ταξινομητών

Example	Features	Label set
1	$\vec{x}_1$	$\{\lambda_1, \lambda_4\}$
2	$\vec{x}_2$	$\{\lambda_3, \lambda_4\}$
3	$\vec{x}_3$	$\{\lambda_1\}$
4	$\vec{x}_4$	$\{\lambda_2, \lambda_3, \lambda_4\}$



Μετατροπή του προβλήματος (3/5)

- Αντιγραφή κάθε παραδείγματος πολλαπλής ετικέτας τόσες φορές όσες και το πλήθος των ετικετών του
- Σε κάθε εμφάνιση το παράδειγμα θα έχει μια ετικέτα από το αρχικό σετ ετικετών του
- Αυξάνεται ο αριθμός των παραδειγμάτων
- Μπορώ να αποδώσω βάρη στα αντιγραμμένα παραδείγματα, ίσα με $1/\text{αριθμό αντιγραφών}$
 - Αυτό απαιτεί την χρήση ταξινομητών που μπορούν να λάβουν υπόψη τους τα βάρη των παραδειγμάτων

ft	Label set		ft	Label	Weight
x ₁	{λ ₁ , λ ₄ }		x ₁	λ ₁	0,50
x ₂	{λ ₃ , λ ₄ }	→	x ₁	λ ₄	0,50
x ₃	{λ ₁ }		x ₂	λ ₃	0,50
x ₄	{λ ₂ , λ ₃ , λ ₄ }		x ₂	λ ₄	0,50
			x ₃	λ ₁	1,00
			x ₄	λ ₂	0,33
			x ₄	λ ₃	0,33
			x ₄	λ ₄	0,33

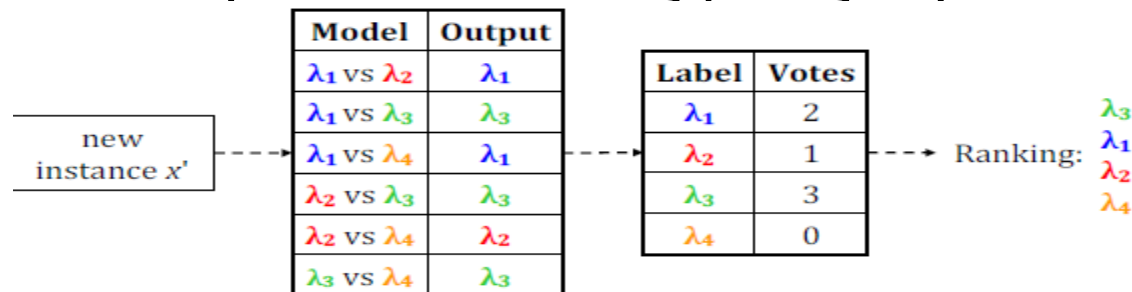
Μετατροπή του προβλήματος (4/5)

- Ανα ζεύγος ετικετών εκπαιδεύεται και ένα μοντέλο που να διακρίνει την μια ετικέτα από την άλλη
- Κάθε μοντέλο εκπαιδεύεται με τα παραδείγματα που έχουν τουλάχιστον μία από τις δυο ετικέτες, αλλά όχι και τις δύο ταυτόχρονα.
- Συνολικά εκπαιδεύονται $q(q-1)/2$ μοντέλα, όπου q το συνολικό πλήθος των ετικετών
- Δεδομένου ενός καινούριου παραδείγματος, πραγματοποιείται κατάταξη των μοντέλων και επιλέγονται οι ετικέτες με τις περισσότερες ψήφους

ft	Label set
x ₁	{λ ₁ , λ ₄ }
x ₂	{λ ₃ , λ ₄ }
x ₃	{λ ₁ }
x ₄	{λ ₂ , λ ₃ , λ ₄ }

ft	λ ₁ vs λ ₂	ft	λ ₁ vs λ ₃	ft	λ ₁ vs λ ₄	ft	λ ₂ vs λ ₃	ft	λ ₂ vs λ ₄	ft	λ ₃ vs λ ₄
x ₁	λ ₁	x ₁	λ ₁	x ₂	λ ₄	x ₂	λ ₃	x ₁	λ ₄	x ₁	λ ₄
x ₃	λ ₁	x ₂	λ ₃	x ₃	λ ₁			x ₂	λ ₄		
x ₄	λ ₂	x ₃	λ ₁	x ₄	λ ₄						
		x ₄	λ ₃								

αποφάσεων των



Μετατροπή του προβλήματος (5/5)

- Συλλογική μάθηση:
- Εκπαιδεύονται q μοντέλα, ένα σε κάθε επανάληψη, ένα για κάθε (δυναμική) ετικέτα.
- Το σετ των χαρακτηριστικών εισόδου επεκτείνεται και περιλαμβάνει και τις ετικέτες των προηγ. μοντέλων.
- Στην πρόβλεψη, οι μεμονωμένες αποφάσεις μεταδίδονται, εμπλουτίζοντας τον χώρο των χαρακτηριστικών

ft	Label set
x ₁	{λ ₁ , λ ₄ }
x ₂	{λ ₃ , λ ₄ }
x ₃	{λ ₁ }
x ₄	{λ ₂ , λ ₃ , λ ₄ }

ft	λ ₁
x ₁	true
x ₂	false
x ₃	true
x ₄	false

ft, λ ₁	λ ₂
x ₁ , true	false
x ₂ , false	false
x ₃ , true	false
x ₄ , false	true

ft, λ ₁ , λ ₂	λ ₃
x ₁ , true, false	false
x ₂ , false, false	true
x ₃ , true, false	false
x ₄ , false, true	true

ft, λ ₁ , λ ₂ , λ ₃	λ ₄
x ₁ , true, false, false	true
x ₂ , false, false, true	true
x ₃ , true, false, false	false
x ₄ , false, true, true	true

Μετατροπή του προβλήματος (5/5)

- Συλλογική μάθηση:
- Εκπαιδεύονται q μοντέλα, ένα σε κάθε επανάληψη, ένα για κάθε (δυναμική) ετικέτα.
- Το σετ των χαρακτηριστικών εισόδου επεκτείνεται και περιλαμβάνει και τις ετικέτες των προηγ. μοντέλων.
- Στην πρόβλεψη, οι μεμονωμένες αποφάσεις μεταδίδονται, εμπλουτίζοντας τον χώρο των χαρακτηριστικών

ft	Label set
x ₁	{λ ₁ , λ ₄ }
x ₂	{λ ₃ , λ ₄ }
x ₃	{λ ₁ }
x ₄	{λ ₂ , λ ₃ , λ ₄ }

ft	λ ₁
x ₁	true
x ₂	false
x ₃	true
x ₄	false

ft, λ ₁	λ ₂
x ₁ , true	false
x ₂ , false	false
x ₃ , true	false
x ₄ , false	true

ft, λ ₁ , λ ₂	λ ₃
x ₁ , true, false	false
x ₂ , false, false	true
x ₃ , true, false	false
x ₄ , false, true	true

ft, λ ₁ , λ ₂ , λ ₃	λ ₄
x ₁ , true, false, false	true
x ₂ , false, false, true	true
x ₃ , true, false, false	false
x ₄ , false, true, true	true

Προσαρμογή του αλγορίθμου

- C4.5 πολλαπλών ετικετών (Multi-label C4.5)

- Κλασσική εντροπία (όταν έχω μια ετικέτα):

$$entropy(S) = - \sum_{i=1}^N p(c_i) \log p(c_i)$$

- N το πλήθος των τιμών της ετικέτας
- $p(c_i)$ η πιθανότητα η ετικέτα μου να πάρει την τιμή c_i .
- όταν έχω πολλαπλές ετικέτες:

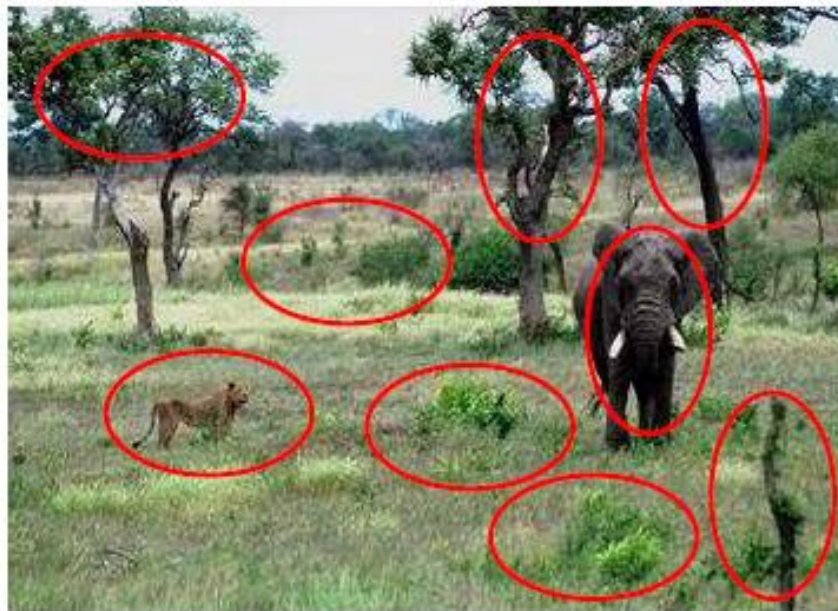
$$- \sum_{i=1}^N ((p(c_i) \log p(c_i)) + (q(c_i) \log q(c_i)))$$

- όπου $p(c_i)$ η πιθανότητα η ετικέτα μου να πάρει την τιμή c_i .
- και $q(c_i) = 1 - p(c_i)$ η πιθανότητα η ετικέτα μου να μην πάρει την τιμή c_i .

Μάθηση πολλαπλών παραδειγμάτων και πολλαπλών ετικετών

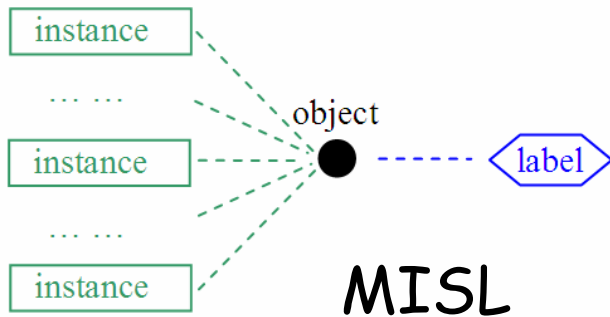
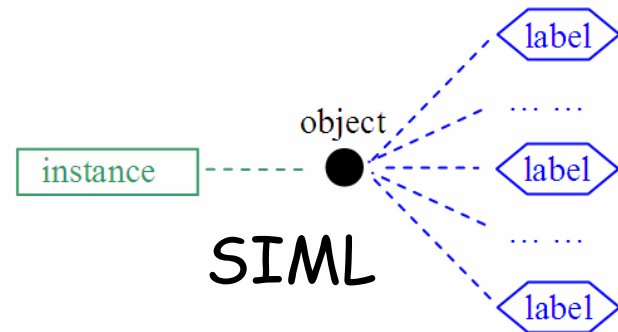
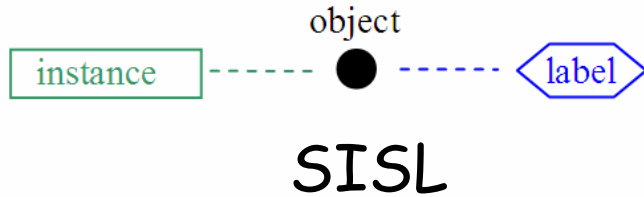
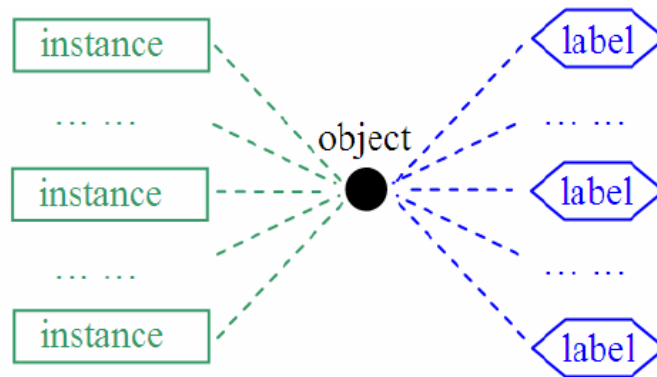
Multi-instance & multi-label Learning

- Μια εικόνα συχνά περιέχει πολλαπλές περιοχές, και χρειάζεται διαφορετικό παράδειγμα για την αναπαράσταση της κάθε περιοχής.
- Ταυτόχρονα, η εικόνα μπορεί να ανήκει σε περισσότερες ετικέτες



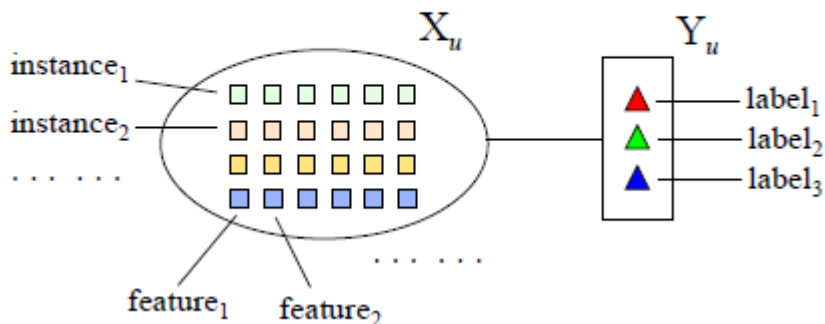
Elephant
Lion
Grassland
Tropic
Africa
... ..

MIML Learning

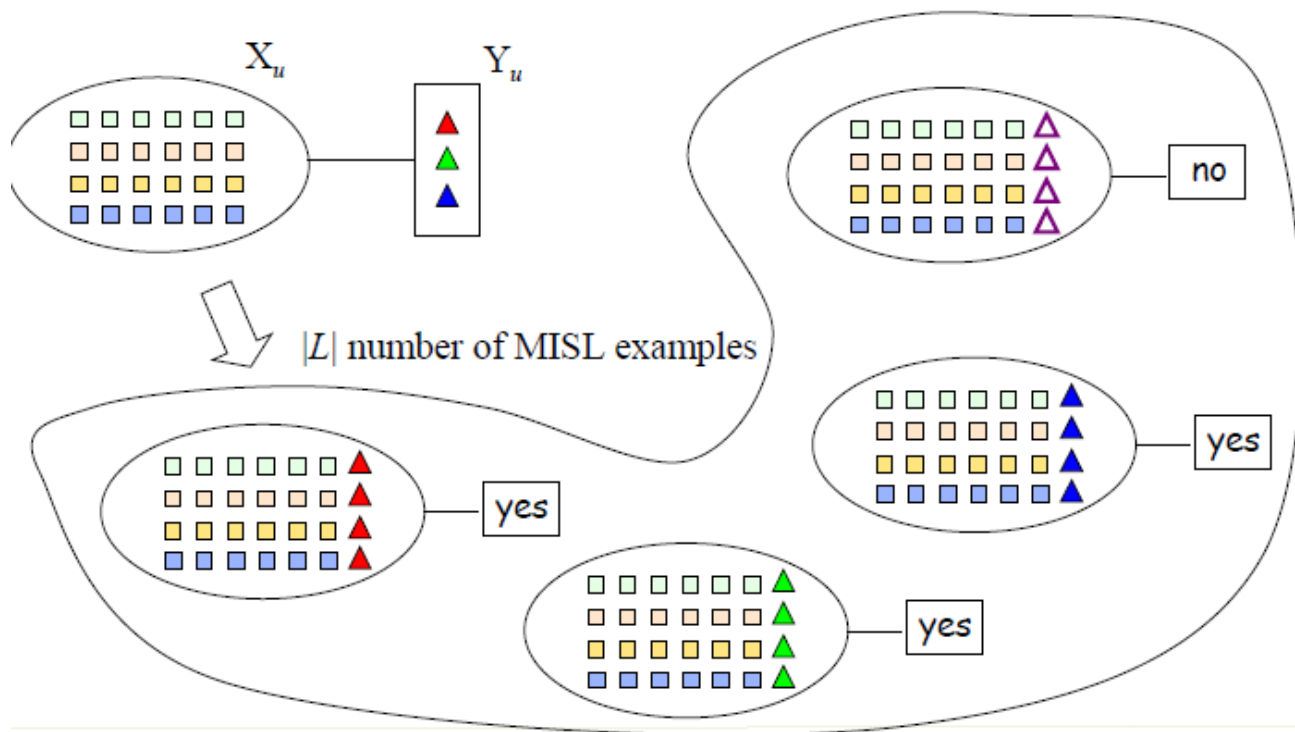


Μετατροπή σε Multi-instance

An MIML example (X_u, Y_u)

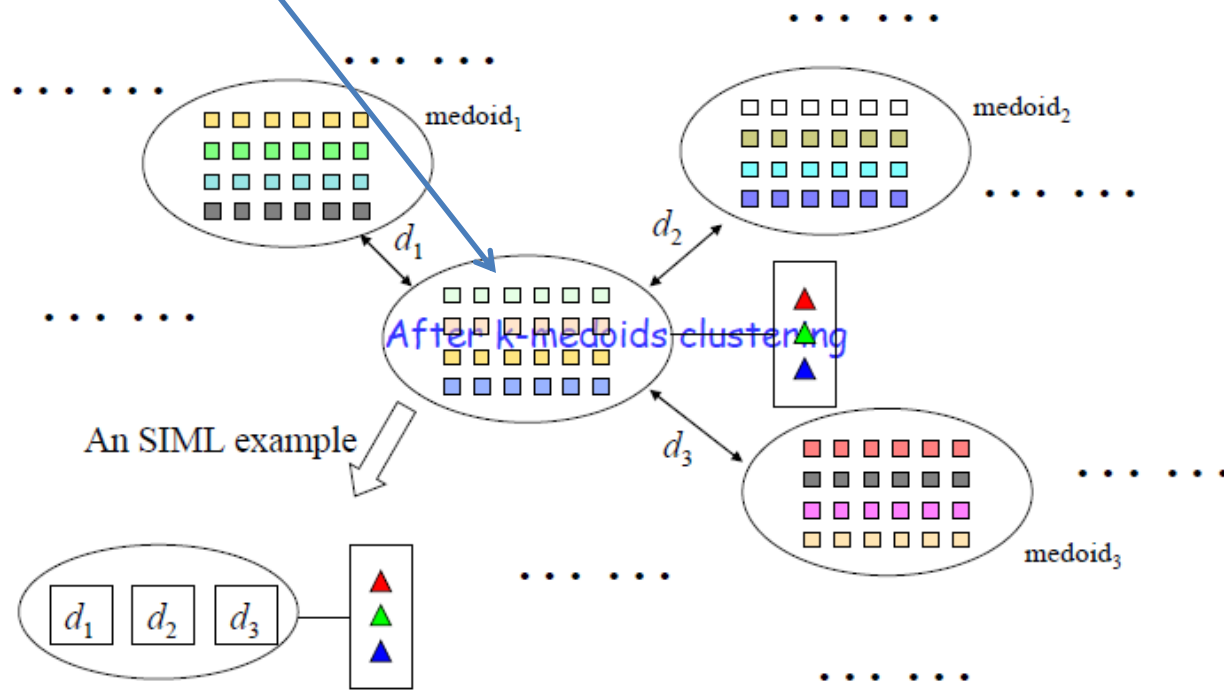
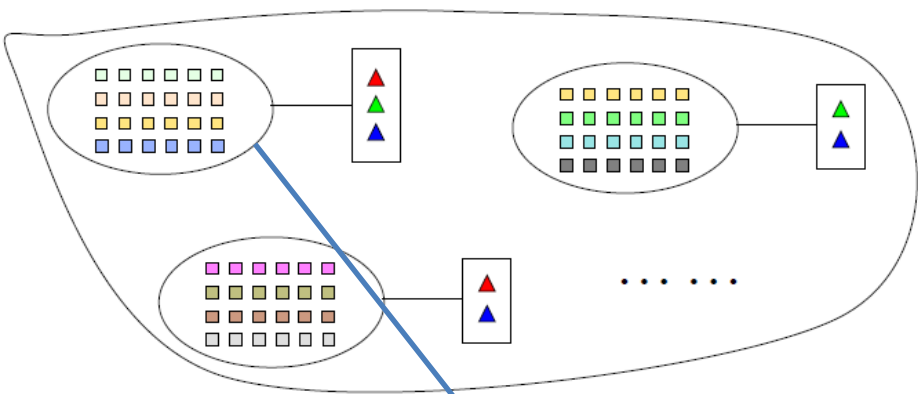


Label set L



Μετατροπή σε Multi-Label

Σε κάθε σακούλα παραδειγμάτων εφαρμόζεται ομαδοποίηση (πχ k-medoids clustering). Τα k κέντρα είναι τα χαρακτηριστικά του καινούριου μοναδικού παραδείγματος



Βιβλιογραφία

- <http://www.citeulike.org/group/7105/tag/multilabel>
- <http://mulan.sourceforge.net>
- <http://people.oregonstate.edu/~sorowerm/pdf/Qual-Multilabel-Shahed-CompleteVersion.pdf>
- <http://papers.nips.cc/paper/3047-multi-instance-multi-label-learning-with-application-to-scene-classification.pdf>