Διαγωνισμός Yelp στην πλατφόρμα Kaggle

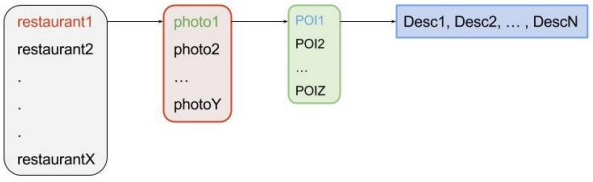
Αντιγόνη Μ. Φούντα
ΑΕΜ: 647
founanti@csd.auth.gr

Θωμάς Παπαγεωργίου
ΑΕΜ: 639
ppthomas@csd.auth.gr

ΠΕΡΙΛΗΨΗ
Η παρούσα αναφορά αποτελεί την περιγραφή της μεθοδολογίας η οποία χρησιμοποιήθηκε για την συμμετοχή μας στον διαγωνισμό του Yelp στην πλατφόρμα Kaggle. Η τελική ακρίβεια των αποτελεσμάτων στα άγνωστα δεδομένα έφτασε το 50,1 % παίρνοντας υπόψιν όλα τα εστιατόρια του training και του test set και ένα μικρό δείγμα εικόνων για κάθε εστιατόριο.

Λέξεις Κλειδιά
Multi-instance, multi-label, yelp, kaggle.

1. ΕΙΣΑΓΩΓΗ
Σκοπός της εργασίας είναι η ανάπτυξη ενός συστήματος σημασιολογικού χαρακτηρισμού εστιατορίων, σύμφωνα με κάποια σύνολα δεδομένων της εφαρμογής Yelp, για τον αντίστοιχο διαγωνισμό[1] της πλατφόρμας Kaggle. Τα δεδομένα αποτελούνται από εστιατόρια, κάθε ένα εκ των οποίων αποτελείται από ένα πλήθος φωτογραφιών. Το πρόβλημα που δημιουργείται ως προς την διαχείριση αυτού του πειράματος είναι ότι υπάρχουν πολλά χαρακτηριστικά που περιγράφουν τις εικόνες από κάθε εστιατόριο (Εικ. 1.1), επομένως αντιμετωπίζουμε μάθηση από πολλαπλές περιπτώσεις (multi-instance learning), ενώ κάθε εστιατόριο πρέπει να χαρακτηριστεί για ένα πλήθος ετικετών, επομένως έχουμε και μάθηση από δεδομένα πολλαπλών ετικετών (multi-label learning).



Εικόνα 1.1. Αναπαράσταση των δεδομένων πολλαπλών περιπτώσεων του dataset.

Δίνονται δεδομένα εκπαίδευσης (training set) που αφορούν 2001 εστιατόρια και περίπου 235.000 εικόνες των εστιατορίων, καθώς και τον χαρακτηρισμό των εστιατορίων αυτών ως προς εννέα κατηγορίες, όπως αναφέρθηκε παραπάνω. Στην συνέχεια υπάρχει ένα σύνολο από 10.001 εστιατόρια με τις αντίστοιχες, περίπου 240.000, φωτογραφίες τους (test set), τα οποία θα πρέπει να αντιστοιχηθούν με κάποιες από τις εννιά κατηγορίες. Οι κατηγορίες σύμφωνα με τις οποίες θα πρέπει να χαρακτηριστούν τα δεδομένα είναι οι εξής:

1. Κατάλληλο για μεσημεριανό (good_for_lunch)
2. Κατάλληλο για βραδινό (good_for_dinner)
3. Κάνει κρατήσεις (takes_reservations)
4. Έχει καθίσματα έξω (outdoor_seating)
5. Είναι ακριβό (restaurant_is_expensive)
6. Σερβίρει αλκοόλ (has_alcohol)
7. Έχει σερβιτόρους (has_table_service)
8. Έχει κομψό περιβάλλον (ambience_is_classy)
9. Κατάλληλο για παιδιά (good_for_kids)

Για την καλύτερη δυνατή πραγματοποίηση των πειραμάτων η εργασία χωρίστηκε σε δύο μέρη. Στο πρώτο μέρος πραγματοποιήθηκαν δύο βήματα ομαδοποίησης για να βγουν τα διανύσματα που χαρακτηρίζουν κάθε εστιατόριο, ενώ στο δεύτερο μέρος γίνεται εκπαίδευση στο σύνολο δεδομένων που βγήκε από το προηγούμενο μέρος και χαρακτηρίζονται τα εστιατόρια ως προς τις κατηγορίες. Η μεθοδολογία περιγράφεται εκτενέστερα στην ενότητα 3, και τα εργαλεία που χρησιμοποιήθηκαν για να υλοποιηθεί όλο το πείραμα περιγράφονται παρακάτω, στην ενότητα 2. Κατόπιν, παρουσιάζονται στην ενότητα 4 τα αποτελέσματα ως προς την ακρίβεια πρόβλεψης με βάση την αξιολόγηση ως προς το training set αλλά και τα αποτελέσματα από το Kaggle για την αξιολόγηση ως προς το test set.

2. ΕΡΓΑΛΕΙΑ
Για το πρώτο μέρος της εργασίας, την παραγωγή των χαρακτηριστικών για κάθε εστιατόριο, χρησιμοποιήσαμε την βιβλιοθήκη μηχανικής όρασης openCV[1], μέσα από την γλώσσα Python, για να βγάλουμε σημεία ενδιαφέροντος από τις εικόνες. Πιο συγκεκριμένα, για την εξαγωγή των σημείων ενδιαφέροντος (POI - points of interest) χρησιμοποιήσαμε την τεχνική εξαγωγής χαρακτηριστικών SURF[2], η οποία χρησιμοποιείται για Αναγνώριση Αντικειμένων (Object Recognition) και βασίζεται στην τεχνική SIFT[3], αλλά είναι αρκετές φορές πιο γρήγορη. Κατόπιν, χρησιμοποιήθηκε ο αλγόριθμος k-means της βιβλιοθήκης scikit-learn για την διαδικασία του clustering και παράχθηκε ο συντελεστής σιλουέτας για κάθε ένα από τα δύο βήματα ομαδοποίησης, για να υπολογίσουμε την απόδοση της ομαδοποίησης. Τέλος, για το πρώτο μέρος της εργασίας χρησιμοποιήθηκαν και οι βιβλιοθήκες της Python, Numpy & CSV.

Στο δεύτερο μέρος της εργασίας, για τον χαρακτηρισμό των εστιατορίων του test set, χρησιμοποιήσαμε τη βιβλιοθήκη μάθησης από πολλαπλές ετικέτες MULAN[4], μέσα από την γλώσσα Java, η οποία εκπαιδεύει δύο μοντέλα μάθησης για classification, τον Multi-label kNN [5] και τον RAkEL (random k-labelsets) [6]. Η έκδοση της Mulan που χρησιμοποιήθηκε είναι η 1.5, για την οποία απαραίτητη προϋπόθεση για να εκτελεστεί είναι να υπάρχει στο project η βιβλιοθήκη του Weka 3.7.10 και το Junit 4.10.

Όλος ο κώδικας της υλοποίησης βρίσκεται στο παρακάτω repository στο GitHub: https://github.com/AndyFou/yelp-contest.

[1] https://www.kaggle.com/c/yelp-restaurant-photo-classification

3. ΜΕΘΟΔΟΛΟΓΙΑ

Η μεθοδολογία που ακολουθήσαμε για να υλοποιήσουμε το συγκεκριμένο πρόβλημα χωρίζεται σε 2 μέρη· στο πρώτο μέρος πραγματοποιούμε δύο φάσεις ομαδοποίησης με τον k-means για να μετατρέψουμε το πρόβλημα του Multi-instance learning σε Single-instance και στο δεύτερο μέρος εκτελούμε δύο αλγορίθμους ταξινόμησης (MLkNN και RAkEL) για να χαρακτηρίσουμε τα δεδομένα του test set.

```
PSEUDOCODE FOR CLUSTERING & CLASSIFICATION

//Get POI from photos
Load data
POIvector = SURF(data)      //get features of POI

//1st Clustering
numclusters=4096
kmeans(POIvector,numclusters)    //perform clustering on POIvector
PhotoVector = Assign clusters    //assign clusters to create Photovector

//2nd Clustering
numclusters=100
kmeans(PhotoVector,numclusters)  //perform clustering on POIvector
RestVector = Assign clusters     //assign clusters to create restaurant
vector

labeled data = RestVector + labels    //Create training dataset

//Classification
model = train RAkEL classifier(labeled data) //train classification model
model.predict(unlabeled data)    //Predict labels for test set
```

Εικόνα 3.1. Η μεθοδολογία που ακολουθήσαμε σε μορφή ψευδοκώδικα.

Εξαιτίας του τεράστιου όγκου των δεδομένων και της μικρής υπολογιστικής ισχύος που είχαμε στη διάθεσή μας, για να πραγματοποιηθούν τα πειράματα έγιναν δύο δοκιμές για δειγματοληψία. Στο πρώτο πείραμα πήραμε ένα δείγμα από τα εστιατόρια (50 εστιατόρια) με όσες εικόνες περιείχε το κάθε εστιατόριο (περίπου 13000 εικόνες) ενώ στην δεύτερη περίπτωση πήραμε όλα τα εστιατόρια του training set και έναν μικρό αριθμό εικόνων από κάθε εστιατόριο.

3.1 Μέρος 1ο – Clustering

Στο πρώτο μέρος της εργασίας πραγματοποιείται διπλή ομαδοποίηση για να παραχθεί το τελικό dataset το οποίο θα δοθεί σαν όρισμα στο δεύτερο μέρος για να γίνει το classification. Η διαδικασία που ακολουθείται στο πρώτο μέρος αναπαρίσταται γραφικά στην Εικόνα 3.2. Όπως φαίνεται στην εικόνα, αρχικά ξεκινάμε με ένα δείγμα εικόνων από το οποίο, για κάθε εικόνα, εξάγονται σημεία ενδιαφέροντος μέσω του SURF. Εφόσον υπολογιστούν οι descriptors, δημιουργείται το διάνυσμα κάθε σημείου ενδιαφέροντος (POIvector) το οποίο αποτελείται από έναν σταθερό αριθμό 64 descriptors. Κάθε εικόνα αποτελείται από ένα πλήθος POIvectors, το οποίο είναι μεταβλητού μεγέθους για κάθε εικόνα και εξαρτάται από τα σημεία ενδιαφέροντος που προέκυψαν από το SURF.

Κατόπιν πραγματοποιείται η πρώτη φάση του clustering. Για να γίνει αυτό μετατρέψαμε το μεταβλητό πλήθος των POIvectors σε σταθερό, κρατώντας την μικρότερη τιμή που υπήρξε σε εικόνα, και στην συνέχεια εκτελείται η ομαδοποίηση με τον αλγόριθμο k-means και 4096 clusters. Το αποτέλεσμα που παράγεται για κάθε vector προσμετράτε αθροιστικά και δημιουργείται για κάθε εικόνα ένα διάνυσμα το οποίο αποτελείται από τις συχνότητες εμφάνισης κάθε cluster (από 0 μέχρι όσο ήταν το πλήθος των POI). Αυτό ονομάζεται διάνυσμα εικόνας (PhotoVector).

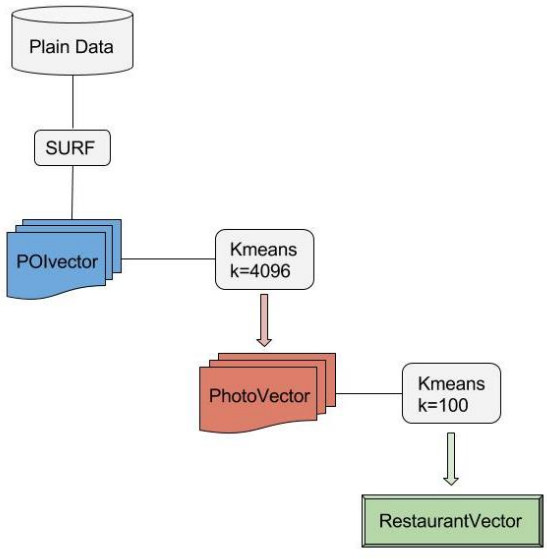
Τέλος, πραγματοποιείται η δεύτερη φάση του clustering, για να βγουν τα τελικά διανύσματα των εστιατορίων, δηλαδή το τελικό dataset. Κάθε εστιατόριο περιέχει πολλές εικόνες, επομένως κάθε εστιατόριο χαρακτηρίζεται από τόσα PhotoVectors όσα το πλήθος των εικόνων που περιέχει. Εκτελείται ξανά ο αλγόριθμος k-means, αυτή τη φορά με 100 κέντρα, και το αποτέλεσμα που παράγεται επίσης προσμετράτε αθροιστικά ώστε να δημιουργηθεί ένα διάνυσμα το οποίο θα έχει για κάθε εστιατόριο μία τιμή συχνότητας για κάθε cluster. Αυτό, σε συνδυασμό με τις ετικέτες κάθε εστιατορίου του training set που δίνονται από το Kaggle, αποτελεί το τελικό μας σύνολο δεδομένων το οποίο θα εκπαιδεύσουμε στο δεύτερο μέρος της εργασίας ώστε να προβλέψουμε τις ετικέτες για τα «άγνωστα» εστιατόρια (αυτά του test set).

Ο αριθμός των ομάδων που δίνεται ως είσοδος στον k-means για κάθε φάση υπολογίστηκε εμπειρικά. Για την πρώτη φάση της ομαδοποίησης χρησιμοποιήθηκε και η μετρική του Συντελεστή Σιλουέτας, για να συμπεράνουμε εάν το clustering που πραγματοποιήθηκε είχε καλά αποτελέσματα, οπότε να συγκρίνουμε εν τέλη τους αριθμούς των ομάδων που ζητήσαμε. Έγιναν πειράματα για τρία διαφορετικά k, και τα αποτελέσματα του συντελεστή φαίνονται στον Πίνακα 3.1. Λόγω των αποτελεσμάτων αυτών, για την πρώτη φάση επιλέχθηκε ο αριθμός k=4096.

Το τελικό dataset που προκύπτει από το πρώτο μέρος είναι σε μορφή .arff, δηλαδή σε τύπο αρχείου που υποστηρίζει το Weka, και αποτελείται από το id του εστιατορίου, την συχνότητα εμφάνισης για κάθε cluster και μία δυαδική τιμή για κάθε ετικέτα, αν δηλαδή η κάθε ετικέτα χαρακτηρίζει αυτό το εστιατόριο ή όχι.

	k=1024	**k=2048**	**k=4096**
1st Round	0,16510	0,0941	0,4209

Πίνακας 3.1. Αποτελέσματα Συντελεστή Σιλουέτας για τον πρώτο γύρο του clustering.



Εικόνα 3.2. Αποτελέσματα του training set

3.2 Μέρος 2ο – Classification

Στο δεύτερο μέρος της εργασίας πραγματοποιείται ο χαρακτηρισμός των δεδομένων του test set με βάση τις 9 ετικέτες που αναφέρθηκαν στην εισαγωγή. Πιο συγκεκριμένα, γίνεται εκπαίδευση δύο μοντέλων classification με τους αλγορίθμους RAkEL και MLkNN με το dataset που βγήκε από το πρώτο μέρος, και στην συνέχεια δίνεται το άγνωστο σύνολο δεδομένων για να χαρακτηριστεί. Με βάση το μοντέλο, τα αποτελέσματα που βγήκαν έδιναν τις τιμές true ή false για κάθε label σε κάθε εστιατόριο. Επιπλέον από τους αλγορίθμους δίνεται και μία κατάταξη (ranking) των labels αλλά αυτό δεν χρησιμοποιήθηκε από εμάς στην τελική πρόβλεψη.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

Όπως αναφέρθηκε και στην προηγούμενη ενότητα, τα πειράματα που πραγματοποιήθηκαν ακολουθούν δύο προσεγγίσεις δειγματοληψίας. Στην πρώτη περίπτωση πήραμε τα 50 πρώτα εστιατόρια με όλες τις φωτογραφίες τους ενώ στην δεύτερη περίπτωση έγιναν δοκιμές με όλο το σύνολο των εστιατορίων και τις τέσσερις πρώτες φωτογραφίες από κάθε εστιατόριο.

Στον Πίνακα 4.1 αναφέρονται τα αποτελέσματα της αξιολόγησης των δύο αλγορίθμων classification για το δείγμα των εστιατορίων του training set, με 10-fold cross validation, ενώ στον Πίνακα 4.2 αναφέρονται τα αποτελέσματα της αξιολόγησης για το σύνολο των εστιατορίων. Τα clusters που χαρακτηρίζουν το τελικό διάνυσμα κάθε εστιατορίου είναι 100.

	RAkEL	**MLkNN**
Precision	0,5725±0,1094	0,5843±0,1183
Recall	0,6659±0,0894	0,5899±0,1546
F-Measure	0,5744±0,0912	0,5421±0,0818
Accuracy	0,4396±0,0968	0,3964±0,0693
Hamming Loss	0,4125±0,0970	0,4225±0,0518

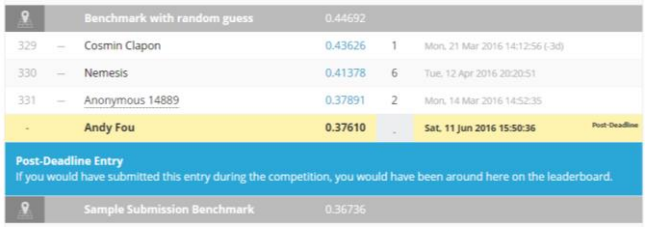
Πίνακας 4.1. Τα Example-Based αποτελέσματα της αξιολόγησης κάθε αλγορίθμου για το δείγμα εστιατορίων του training set.

	RAkEL	**MLkNN**
Precision	0,5762±0,0127	0,5978±0,0154
Recall	0,5192±0,0202	0,6303±0,0335
F-Measure	0,5016±0,0158	0,5753±0,0170
Accuracy	0,3789±0,0151	0,4297±0,0171
Hamming Loss	0,4243±0,0136	0,3901±0,0120

Πίνακας 4.2. Τα Example-Based αποτελέσματα της αξιολόγησης κάθε αλγορίθμου για όλα τα εστιατόρια του training set.

Από τους πίνακες παρατηρούμε ότι τα καλύτερα αποτελέσματα παράγονται με τον αλγόριθμο MLkNN και όλα τα εστιατόρια, καθώς με αυτόν τον συνδυασμό υπάρχει χαμηλότερο hamming loss, δηλαδή «χάσιμο» πληροφορίας, ενώ και το F-Measure έχει την υψηλότερη τιμή. Το accuracy δεν είναι το ψηλότερο που πήραμε αλλά είναι πολύ κοντά στην μεγαλύτερη τιμή (η οποία βγήκε από τον συνδυασμό του αλγορίθμου RAkEL και του δείγματος εστιατορίων), επομένως θεωρούμε ότι η διαφορά δεν είναι σημαντική.

Στην συνέχεια δοκιμάσαμε να κάνουμε πρόβλεψη των άγνωστων δεδομένων και με τις δύο προσεγγίσεις και εισάγαμε τα αποτελέσματα στο Kaggle. Στις Εικόνες 4.1 και 4.2 φαίνονται τα αποτελέσματα του Kaggle για την πρώτη περίπτωση, του δείγματος, και για την δεύτερη, με όλα τα εστιατόρια, αντίστοιχα. Όπως είναι εμφανές, στην δεύτερη περίπτωση τα αποτελέσματα βελτιώθηκαν σημαντικά ανεβάζοντας το τελικό accuracy κατά 13%.



Εικόνα 4.1. Τα αποτελέσματα του Kaggle για την προσέγγιση με το δείγμα των 50 εστιατορίων.

Εικόνα 4.2. Τα αποτελέσματα του Kaggle για την προσέγγιση με το σύνολο όλων των εστιατορίων.

5. ΑΝΑΦΟΡΕΣ

[1] OpenCV: http://opencv.org/

[2] Bay, H, Tuytelaars, T., & Van Gool, L., 2006. SURF: Speeded up Robust Features. In *Computer vision–ECCV 2006* (May 7), 404-417. Springer Berlin Heidelberg.

[3] Lowe, D.G., 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999*. The proceedings of the seventh IEEE international conference on (Vol. 2, pp. 1150-1157). IEEE.

[4] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J. & Vlahavas, I., 2011. Mulan: A java library for multi-label learning. *The Journal of Machine Learning Research*, 12, 2411-2414.

[5] Zhang, M.L., Zhou, Z.H., 2007. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recogn.*. 40(7): 2038-2048.

[6] Tsoumakas, G., Katakis, I., Vlahavas, I., 2001. Random k_Labelsets for Multi-Label Classification. *IEEE Transactions on Knowledge and Data Engineering*. 23(7): 1079-1089.