ΠΜΣ ΕΠΙΣΤΗΜΗ ΚΑΙ ΜΗΧΑΝΙΚΗ ΔΕΔΟΜΕΝΩΝ

ΑΥΤΟΜΑΤΗ ΚΑΤΑΤΜΗΣΗ ΚΑΙ ΚΑΤΗΓΟΡΙΟΠΟΙΣΗ ΤΗΣ ΠΟΛΥΔΙΑΣΤΑΤΗΣ ΧΡΟΝΟΣΕΙΡΑΣ ΑΙΣΘΗΤΗΡΙΑΚΩΝ ΣΗΜΑΤΩΝ

Γεώργιος Βαρδάκας

Υπεύθυνος Καθηγητής:
Παρσόπουλος Κωνσταντίνος

Τμήμα Μηχανικών Η/Υ και Πληροφορικής
Πολυτεχνική Σχολή Πανεπιστημίου Ιωαννίνων

Περιεχόμενα

1 Εισαγωγή

Στην παρούσα εργασία κληθήκαμε να λύσουμε το πρόβλημα της αυτόματη κατάτμηση και κατηγοριοποίησης της πολυδιάστατης χρονοσειράς αισθητηριακών σημάτων. Για να το πετύχουμε αυτό αρχικά συλλέχθηκαν τα δεδομένα που είναι απαραίτητα για την κατασκευή ενός μαθηματικού μοντέλου μηχανικής μάθησης, και στην συνέχεια ακολούθησε η προ-επεξεργασία τους, ώστε να είναι κατάλληλα προετοιμασμένα για να εισαχθούν στο μοντέλο. Επιπρόσθετα μελετήθηκαν διάφορα μοντέλα μηχανικής μάθησης, μέχρι να βρεθεί αυτό με τα καλύτερα αποτελέσματα. Τελικά η αξιολόγηση του μοντέλου μηχανικής μάθησης για τον έλεγχο της ποιότητας των αποτελεσμάτων του πραγματοποιήθηκε με K-fold cross validation.

2 Συλλογή Δεδομένων

Μία προαπαίτηση για την κατασκευή ενός μαθηματικού μοντέλου μηχανικής μάθησης είναι φυσικά τα δεδομένα, καθώς αυτές οι μέθοδοι μαθαίνουν μέσα από αυτά. Για την συλλογή των δεδομένων έγινε χρήση του έξυπνου ρολογιού Fitbit Versa. Το έξυπνο αυτό ρολόι, μας δίνει την δυνατότητα να αντλήσουμε τις μετρήσεις που καταγράφουν οι διαθέσιμοι αισθητήρες του. Οι αισθητήρες αυτοί αποτελούνται από από το γυροσκόπιο, το επιταχυνσιόμετρο καθώς και τον αισθητήρα καρδιακών παλμών. Οι αισθητήρες του γυροσκοπίου και του επιταχυνσιομέτρου καταγράφουν μετρήσεις με συχνότητα 10 Η ενώ ο αισθητήρας καρδιακών παλμών καταγράφει μετρήσεις με συχνότητα 1 Η. Στην παρούσα εργασία έγινε χρήση μόνο των αισθητήρων του γυροσκοπίου και του επιταχυνσιομέτρου καθώς οι καταγραφές του αισθητήρα των καρδιακών παλμών δεν περιείχε χρήσιμη πληροφορία για την αυτόματη κατηγοριοποίηση της πολυδιάστατης χρονοσειράς των αισθητήρων. Συνολικά καταφέραμε να συλλέξαμε δεδομένα από εφτά διαφορετικούς χρήστες. Τα δεδομένα είναι μορφής πίνακα όπως ο παρακάτω:

	TIMESTAMP	ACCEL_X	ACCEL_Y	ACCEL_Z	GYRO_X	GYRO_Y	GYRO_Z	ACTIVITY_ID	USER_ID
0	2020-08-25 16:23:14.590	-1.230657	5.430215	7.670058	-1.084472	0.181100	-0.076701	107	0
1	2020-08-25 16:23:14.690	-1.743032	4.716721	8.117788	-1.465848	0.301478	-0.136357	107	0
2	2020-08-25 16:23:14.790	-2.229070	4.455745	8.960572	-1.877052	0.394159	-0.137423	107	0
3	2020-08-25 16:23:14.890	-2.496031	3.766193	9.097046	-1.774784	0.359005	-0.070309	107	0
4	2020-08-25 16:23:14.990	-2.711515	2.225478	9.026415	-2.076263	0.344090	-0.100138	107	0

Σχήμα 1: Πρώτες πέντε εγγραφές του πίνακας δεδομένων.

όπου η κολόνα TIMESTAMP αναφέρεται στο χρόνο δειγματοληψίας της εγγραφής, η κολόνα ACCEL_{X,Y,Z} αναφέρεται στην μέτρηση του επιταχυνσιομέτρου στον άξονα $\{x, y, z\}$, η κολόνα GYRO_{X,Y,Z} αναφέρεται στην μέτρηση του γυροσκοπίου στον άξονα $\{x, y, z\}$, η κολόνα ACTIVITY_ID δηλώνει την δραστηριότητα που πραγματοποιεί ο χρήστης και τέλος η κολόνα USER_ID αναφέρεται στον ποίος χρήστης έκανε την δραστηριότητα. Οι μετρήσεις των αισθητήρων καταγράφουν δείγματα στον τρισδιάστατο χώρο, αυτός είναι και ο λόγος που κάθε αισθητήρας έχει μετρήσεις τριών τυχαίων μεταβλητών (x, y, z). Επίσης ο κάθε χρήστης έχει μοναδικό ID. Πλέον τα δεδομένα έχουν κατασκευαστεί και αποθηκευτεί με τέτοιο τρόπο, ώστε να είναι έτοιμα για να χρησιμοποιηθούν για το πρόβλημα της αυτόματης ταξηνόμησης ακολουθίας $\{(X_i, y_i)\}_{i=1}^{N}$, με κατηγορία y_i για κάθε ακολουθία X_i. Η κάθε ακολουθία X_i μοντελοποιείται σαν πολυδιάστατη χρονοσειρά αισθητηριακών σημάτων, έχοντας T_i δείγματα $\langle x_1, x_2, ..., x_{T_i} \rangle_i$, με κατηγορία δραστηριότητας y_i. Τέλος το κάθε δείγμα $x_j = [ACCEL\_X, ACCEL\_Y, ACCEL\_Z, GYRO\_X, GYRO\_Y, GYRO\_Z]_j$ και με $y_j = ACTIVITY\_ID$. Για την διαχείρηση των δεδομένων έγινε χρήση της βιβλιοθήκης pandas [1].

3 Αυτόματη Κατάτμηση

Μετά την συλλογή των δεδομένων που μοντελοποιούνται σαν πολυδιάστατη χρονοσειρά αισθητηριακών σημάτων ακολουθεί η προ-επεξεργασία τους. Το πρώτο βασικό βήμα της προ-επεξεργασίας είναι η αυτόματη κατάτμηση του σήματος. Για την επίτευξη της αυτόματης κατάτμησης έγινε χρήση της συνάρτησης Segment(width, overlap) της βιβλιοθήκης seglearn [2] . Η συνάρτηση λαμβάνει ως είσοδο το αρχικό σήμα και το τμηματοποιεί σε τμήματα μεγέθους width, κάνοντας χρήση ενός επικαλυπτόμενου κυλιόμενου παραθύρου (sliding

window) σταθερού μήκος, στην περίπτωσή μας το width ίσο με 5 δευτερόλεπτα. Με αυτήν την μέθοδο κατασκευάζουμε ένα τρισδιάστατο χρονικό τένσορα $\phi_i = \langle W_1, ..., W_M \rangle$ και για κάθε χρονοσειρά X_i. Ο τένσορας ϕ_i έχει σχήμα $(M_i, width, 6)$ με $width$ το μήκος του παραθύρου και M_i ο αριθμός των παραθύρων που κατασκευάστηκαν για κάθε χρονοσειρά X_i. Το τελικό σύνολο δεδομένων είναι το σύνολο όλων των τμημάτων που παρήχθησαν από το κυλιόμενο παράθυρο και συμβολίζεται $\{W_i, y_i\}_{i=1}^{N_w}$, όπου το N_w είναι το πλήθος των τμημάτων του συνόλου δεδομένων. Το πρόβλημα μηχανικής μάθησης της ταξινόμησης των πολυδιάστατων χρονοσειρών του έργου πλέον διατυπώθηκε και αξιολογήθηκε ως ταξινόμηση του τμηματοποιημένου συνόλου δεδομένων $\{W_i, y_i\}_{i=1}^{N_w}$.

4 Εξαγωγή Χαρακτηριστικών

Το επόμενο βήμα είναι η εξαγωγή των χαρακτηριστικών $F = \mathcal{F}(W)$ του τμηματοποιημένου συνόλου δεδομένων $\{W_i, y_i\}_{i=1}^{N_w}$. Πιο συγκεκριμένα για κάθε τμήμα W_i θα εξαγάγουμε κάποια στατιστικά μεγέθη τα οποία στην συνέχεια θα αποτελέσουν την είσοδο για τον αλγόριθμο της μηχανικής μάθησης. Με αυτήν την ενέργεια από τον χώρο του τμηματοποιημένου σήματος μεταφερόμαστε στον χώρο των χαρακτηριστικών $\mathcal{F} : W \to F$. Αυτή η διαδικασία γίνεται για κάθε τυχαία μεταβλητή του τμήματος W_i ξεχωριστά. Μερικά από τα χαρακτηριστικά (στατιστικά μεγέθη) που χρησιμοποιήθηκαν είναι η μέση τιμή, η διάμεσος, το άθροισμα των τετραγώνων, η τυπική απόκλιση, η διακύμανση, το μέγιστο και ελάχιστο στοιχείο, η λοξότητα, η κύρτωση, η μέση φασματική ενέργεια, το μέσο όρο των απόλυτων τιμών, η ρίζα των μέσων τετραγώνων κα.

5 Προ-επεξεργασία Χαρακτηριστικών

Στην συνέχεια ακολούθησε η κανονικοποίηση των δεδομένων εκπαίδευσης που πλέον είναι τα χαρακτηριστικά που εξαγάγαμε στο προηγούμενο βήμα. Η κανονικοποίηση του συνόλου δεδομένων είναι μια κοινή απαίτηση για πολλούς εκτιμητές μηχανικής μάθησης. Συνήθως αυτό γίνεται αφαιρώντας το μέσο όρο και κλιμακώνοντας τη διακύμανση στη μονάδα. Ωστόσο, οι υπερβολικές τιμές (τα outliers) μπορούν συχνά να επηρεάσουν τη μέση τιμή και την διακύμανση του δείγματος με αρνητικό τρόπο. Για την αντιμετώπιση αυτού το προβλήματος σε αυτό το βήμα διαλέχτηκε η μέθοδος του robust data scaling της βιβλιοθήκης scikit-learn [3] καθώς αγνοεί τις ακραίες τιμές και παρήγαγε τα καλύτερα αποτελέσματα σε σχέση με άλλες μεθόδους κανονικοποίησης που δοκιμάστηκαν. Για να το πετύχει αυτό ο συγκεκριμένος μετασχηματισμός λειτουργεί αφαι-

ρώντας την διάμεσο από του χαρακτηριστικού και στην συνέχεια διαιρώντας ενδοτεταρτημοριακό εύρος του (interquartile range).

6 Το μοντέλο

Πλέον μετά από τα στάδια διαλογής και προ-επεξεργασία των δεδομένων, συνεχίζουμε στον ορισμό του μοντέλου μηχανικής μάθησης με στόχο την αυτόματη κατηγοριοποίηση (ταξινόμηση) της πολυδιάστατης χρονοσειράς των αισθητηριακών σημάτων.

References

[1] T. pandas development team, "pandas-dev/pandas: Pandas," Feb. 2020.

[2] C. W. David Burns, "Seglearn: A python package for learning sequences and time series," *arXiv*, 2018.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.