

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ · ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΕΞΟΥΞΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΑΛΓΟΡΙΘΜΟΙ ΜΑΘΗΣΗΣ

ΕΡΓΑΣΤΗΡΙΑΚΗ ΑΣΚΗΣΗ · 2023–2024

ΠΕΡΙΕΧΟΜΕΝΑ

1	ΕΡΩΤΗΜΑ 1	2
1.1	ΑΝΑΛΥΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ	2
1.2	ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ	3

1 ΕΡΩΤΗΜΑ 1

Για την εισαγωγή και την προεπεξεργασία του `.csv` αρχείου, θα χρησιμοποιήσουμε τις βιβλιοθήκες `pandas` και `matplotlib` της Python.

1.1 ΑΝΑΛΥΣΗ ΣΥΝΟΛΟΥ ΔΕΔΟΜΕΝΩΝ

Εισάγουμε τα `.csv` αρχεία μέσω της `os` βιβλιοθήκης και της `read_csv()` συνάρτησης. Καταρχάς, χρησιμοποιώντας τη `head()` μπορούμε να δούμε τις πρώτες εγγραφές του dataset μας. Για παράδειγμα για το πρώτο αρχείο του συνόλου δεδομένων `S006.csv`:

	timestamp	back_x	back_y	back_z	thigh_x	thigh_y	thigh_z	label
0	2019-01-12 00:00:00.000	-0.760242	0.299570	0.468570	-5.092732	-0.298644	0.709439	6
1	2019-01-12 00:00:00.010	-0.530138	0.281880	0.319987	0.900547	0.286944	0.340309	6
2	2019-01-12 00:00:00.020	-1.170922	0.186353	-0.167010	-0.035442	-0.078423	-0.515212	6
3	2019-01-12 00:00:00.030	-0.648772	0.016579	-0.054284	-1.554248	-0.950978	-0.221140	6
4	2019-01-12 00:00:00.040	-0.355071	-0.051831	-0.113419	-0.547471	0.140903	-0.653782	6

Ίδια μορφολογία παρατηρείται σε όλα τα `.csv` του συνόλου δεδομένων, με κάποιες διαφοροποιήσεις που θα αναλυθούν στη συνέχεια. Μέσω της `info()` παρατηρούμε πώς για κάθε χρονική στιγμή δίνονται οι τιμές των αισθητήρων, αποθηκευμένες ως `float24`, στις τρεις διαστάσεις $[x, y, z]$ για τις περιοχές της πλάτης και του μηρού, καθώς και ένα `int64` label.

Για να ελέγξουμε την ακεραιότητα και να εντοπίσουμε τυχούσες συνέπειες, μέσω της συνάρτησης `concat()` ενώνουμε όλα τα 22 αρχεία σε ένα ενιαίο dataframe. Τρέχοντας την `isnull().sum()` έχουμε:

	sum
timestamp	0
back _x	0
back _y	0
back _z	0
thigh _x	0
thigh _y	0
thigh _z	0
label	0
index	5740689
Unnamed: 0	6323682

Έχουν εμφανιστεί NaN τιμές στις στήλες "index" και "Unnamed: 0", οι οποίες στήλες εμφανίζονται επιπλέον σε κάποιους συμμετέχοντες. Ελέγχοντας όλα τα αρχεία, η στήλη "index" εμφανίζεται στα αρχεία `S015.csv` και `S021.csv` και η στήλη "Unnamed: 0" στο αρχείο `S023.csv`. Έπειτα από έλεγχο, φαίνεται πως πρόκειται για δείκτες αύξουσας αρίθμησης που δεν προσφέρουν κάποια επιπλέον πληροφορία. Επομένως, μπορούμε να τις αφαιρέσουμε χρησιμοποιώντας τη συνάρτηση `drop('όνομα', axis=1)`¹.

Χρησιμοποιώντας τη συνάρτηση `describe()` μπορούμε να υπολογίσουμε βασικές στατιστικές μετρικές για τα δεδομένα μας. Η συνάρτηση επιστρέφει ένα dataframe με τις ακόλουθες στήλες:

- **count**: ο συνολικός αριθμός των μη-μηδενικών τιμών για κάθε στήλη.
- **mean**: ο μέσος όρος των τιμών για κάθε στήλη.
- **min**: η ελάχιστη τιμή για κάθε στήλη.
- **25%**: η τιμή του 25ου εκατοστημορίου για κάθε στήλη.
- **50%**: η τιμή του 50ου εκατοστημορίου για κάθε στήλη.
- **75%**: η τιμή του 75ου εκατοστημορίου για κάθε στήλη.

¹το `axis=1` αναφέρεται στην αφαίρεση στήλης

■ **max**: η μέγιστη τιμή για κάθε στήλη.

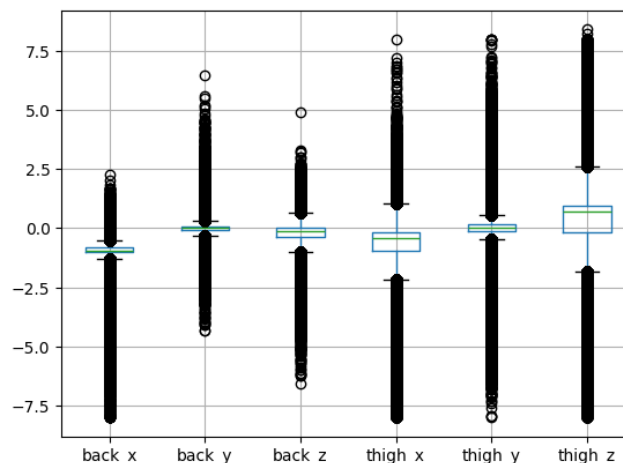
Ενώνοντας συγκεντρωτικά τις μετρήσεις όλων των συμμετεχόντων στο `df_combined` μέσω της `concat()`, αυτά είναι τα βασικά συγκεντρωτικά στατιστικά μεγέθη όπως προκύπτουν από το `describe()` για όλες τις μετρήσεις από τους συμμετέχοντες, έχοντας αφαιρέσει την ετικέτα `label`:

	<code>back_x</code>	<code>back_y</code>	<code>back_z</code>	<code>thigh_x</code>	<code>thigh_y</code>	<code>thigh_z</code>
count	6461328	6461328	6461328	6461328	6461328	6461328
mean	-0.884957	-0.013261	-0.169378	-0.594888	0.020877	0.374916
std	0.377592	0.231171	0.364738	0.626347	0.388451	0.736098
min	-8.000000	-4.307617	-6.574463	-8.000000	-7.997314	-8.000000
25%	-1.002393	-0.083129	-0.372070	-0.974211	-0.100087	-0.155714
50%	-0.974900	0.002594	-0.137451	-0.421731	0.032629	0.700439
75%	-0.812303	0.072510	0.046473	-0.167876	0.154951	0.948675
max	2.291708	6.491943	4.909483	7.999756	7.999756	8.406235

Παρατηρούμε πως οι μετρήσεις των αισθητήρων βρίσκονται στο διάστημα $[-8, 8]$, ενώ η τυπική τους απόκλιση είναι μικρή, το οποίο δείχνει ότι οι μετρήσεις είναι αρκετά συμπυκνωμένες γύρω από τον μέσο όρο που είναι κοντά στο μηδέν. Μπορούν να διεξαχθούν επιπλέον συμπεράσματα ελέγχοντας κάθε συμμετέχοντα ξεχωριστά.

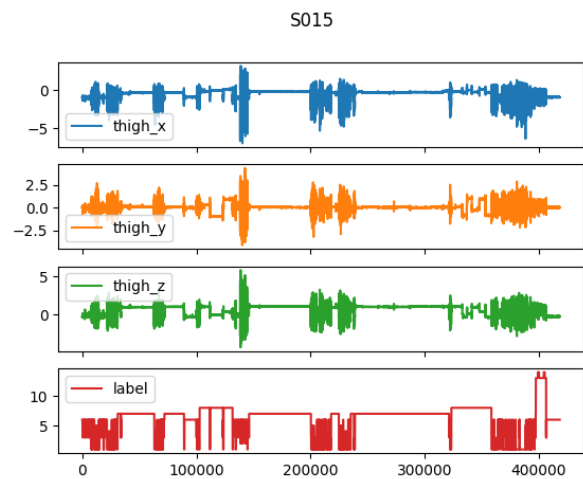
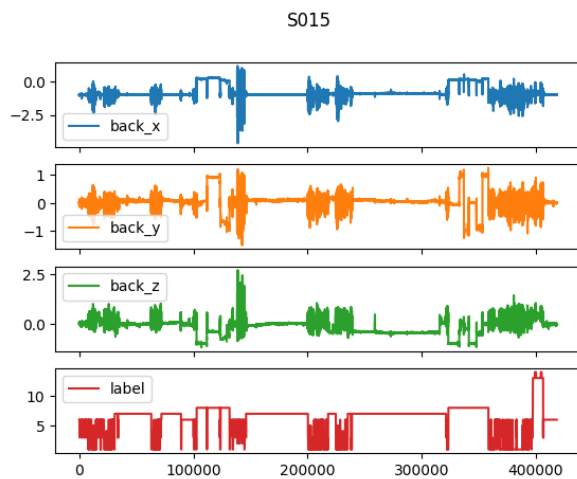
1.2 ΓΡΑΦΙΚΕΣ ΠΑΡΑΣΤΑΣΕΙΣ

Αν χρησιμοποιήσουμε το `plotbox()` της `Matplotlib`, μπορούμε να δημιουργήσουμε το διάγραμμα των τιμών για μια πρώτη οπτικοποίηση των δεδομένων:

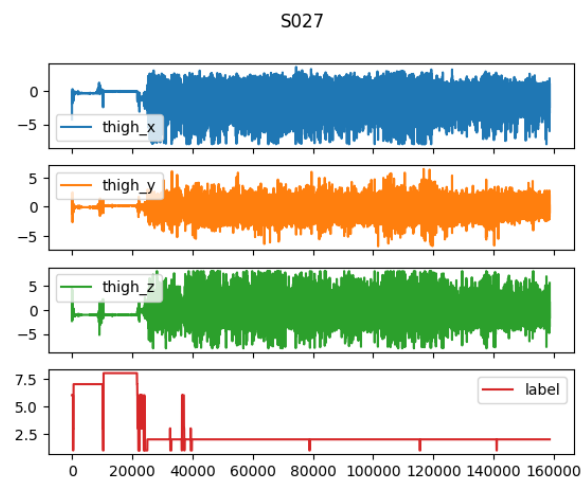
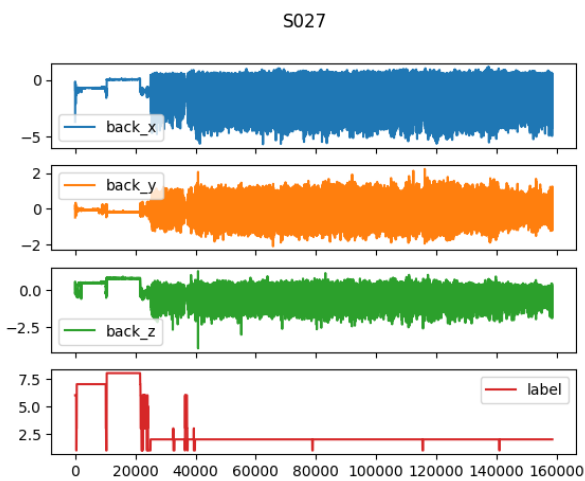


Παρατηρούμε πως επιβεβαιώνονται οι παραπάνω παρατηρήσεις. Επιπλέον φαίνεται πως υπάρχει μια συμμετρικότητα γύρω από το μηδέν για κάθε διάσταση.

Παρακάτω βρίσκονται οι μετρήσεις των αισθητήρων για τον συμμετέχοντα `s015` σε ξεχωριστά subplots για τα δύο επιταχυνσιόμετρα. Είναι προφανές ότι υπάρχει μια συσχέτιση μεταξύ των κινήσεων της πλάτης και του μηρού, καθώς παρατηρούμε ομοιότητες στα διαγράμματα. Επίσης φαίνεται πως υπάρχει μια αντιστοιχία μεταξύ των τιμών της ετικέτας `label` με τις κινήσεις των αισθητήρων.



Παρόλα αυτά δεν παρατηρείται το ίδιο σε όλους του συμμετέχοντες. Για παράδειγμα, στον συμμετέχοντα S027, η *label* δεν έχει μια εμφανή συσχέτιση με τις μετρήσεις των αισθητήρων.



Από τα γραφήματα εξάγεται το συμπέρασμα ότι οι στιγμές με έντονες μεταβολές στις τιμές αντιστοιχούν σε κάποια κίνηση των συμμετεχόντων, ενώ στα σημεία που τα γραφήματα παρουσιάζουν ευθεία γραμμή, οι συμμετέχοντες βρίσκονται σε αδράνεια.

2 ΕΡΩΤΗΜΑ 2

2.1 NEURAL NETWORKS