Τεχνολογία Ήχου και Εικόνας 2018
Παραδοτεό 1

Χριστίνα Θεοδωρίδου - 8055
Φρανκ Μπλάννινγκ - 6698
Αποστόλης Φανάκης - xxxx

2 Νοεμβρίου 2018

Περιεχόμενα

1 Εισαγωγή 2

2 Τι θα κάνουμε 2

3 Γιατί θα το κάνουμε 2

4 Πως το έκαναν άλλοι 2

5 Πως σκοπεύουμε να το κάνουμε εμείς 2

6 Τι άλλο...; 4

1 Εισαγωγή

Το ζητούμενο της εργασίας είναι η ανάπτυξη ενός μοντέλου μηχανικής μάθησης το οποίο παρέχοντας ένα αρχείο ήχου θα μπορεί να ξεχωρίσει ανάμεσα στα κομμάτια του χρόνου που περιέχουν ομιλία (speech) και μουσική (music).

Πρόκειται για ένα πρόβλημα ταξινόμησης που είναι σημαντικό γιατί έχει εφαρμογές σε πλατφόρμες κοινωνικών δικτύων για την αναγνώρισή περιεχομένου με πνευματικά δικαιώματά, σε συστήματα αυτόματης αναγνώρισης διαφημίσεων, μοντέρνα "έξυπνα" βοηθητικά ακοής κ.α. Η πρόσφατη βιβλιογραφία περιέχει θεματολογία όπου στοχεύει είτε στην ανάπτυξή αλγορίθμων για γρήγορη και φθηνή υπολογιστικά ταξινόμηση, είτε στην αναγνώριση πολλής μεγάλης ακρίβειας. Αυτό διότι αυτή τη στιγμή η αναγνώριση με ποσοστό επιτυχίας γύρω στο 98% είναι κάτι συνηθισμένο.

2 Τι θα κάνουμε

3 Γιατί θα το κάνουμε

4 Πως το έκαναν άλλοι

Υπάρχει πληθώρα βιβλιογραφίας σχετική με το θέμα. Έχουν βρεθεί ήδη αρκετές λύσεις, ενώ οι πιο πρόσφατες πετυχαίνουν αξιοσημείωτα αποτελέσματα τόσο όσων αφορά την ταχύτητα του διαχωρισμού όσο και την ακρίβεια των αποτελεσμάτων.

Πιθανές αναφορές:

- ποιά μοντέλα (δέντρα, πιθανοτικά, neural...) είναι αποτελεσματικότερα με βάση τη βιβλιογραφία; Νομίζω ανάλογα με το paper υπάρχουν διαφορετικά αποτελέσματα σχετικά με αυτό (άλλα προτείνουν μπαγεσιανά και άλλα νευρωνικά) άρα παίζει ρόλο η επιλογή των features και στο μοντέλο, να το πούμε αυτό.. Πχ κάποια features έχουν μεγάλο correlation -> τα naive bayes δε τη παλεύουν σε αυτά…
- ποια features χρησιμοποιούνται; Τι σημαίνει το καθένα και πως υπολογίζεται; Πόσο ακριβά είναι υπολογιστικά το καθένα;
- ποια είναι η γενικότερη πορεία που ακολουθείται;
  συνήθως:
  1. παραθυροποίηση (τι τύπου; είναι επικαλυπτόμενα τα παράθυρα; πόσα sec είναι το καθένα;)
  2. feature extraction
  3. μετασχηματισμός του χώρου (βλέπε PCA και άλλες μεθόδους)
  4. training
  5. πρόβλεφη
- Άρα κατά τον σχεδιασμό πρέπει εκτός από τη μέθοδο της παραθυροποίησης, τα features και το μοντέλο να επιλεχθούν επίσης κάποιος μετασχηματισμός (δεν το κάνουν πάντα) ή και άλλες παράμετροι. Τι άλλο preprocessing χρειάζεται;

Σύμφωνα με το paper [1] το back propagation neural network πέτυχε ακρίβεια 89.08%, ενώ το SVM πέτυχε 90.12% και η δική τους υλοποίηση SVM (με τον αλγόριθμο cuckoo), CS-SVM, πέτυχε 92.75%.

5 Πως σκοπεύουμε να το κάνουμε εμείς

Πλάνο επίθεσης

Μετά από μελέτη των προηγούμενων υλοποιήσεων και πειραματισμό με την εξαγωγή διάφορων χαρακτηριστικών (features) [και καλά ;)] κλπ κλπ αποφασίσαμε να ακολουθήσουμε την παρακάτω πορεία αντιμετώπισης του προβλήματος:

1. Τι παραθυροποίηση θα κάνουμε (λογικά την κλασική hamming... φαίνεται να είναι φιλοστάνταρ)

2. Ποια features σκοπεύουμε να χρησιμοποιήσουμε; (γιατί;) Χαρακτηριστικά από το πεδίο του χρόνου, το πεδίο της συχνότητας, το cepstral πεδίο, άλλα…
3. Γενικά MFCC + MPEG-7 audio descriptors + ίσως κάνα δυο ακόμα είναι υπέρ-αρκετά
4. Τι μοντέλο/μοντέλα θα δοκιμάσουμε; (γιατί;)
5. Stack (python/R, τι βιβλιοθήκες/kits για τα μοντέλα;)

Διάφορα features από τη βιβλιογραφία:
M. Kashif Saeed Khan · Wasfi G. Al-Khatib Machine-learning based classification of speech and music

1. Percentage of low energy frames
2. Roll off point
3. Spectral flux
4. Zero crossing rate
5. Spectral centroid
6. 4Hz modulation energy
7. Variance of the roll off point
8. Variance of the spectral centroid
9. Variance of the spectral flux
10. Cepstral residual
11. Variance of the cepstral residual

1. Cepstral coefficients
2. Delta cepstral coefficients
3. Harmonic coefficients
4. 4 Hz harmonic coefficients
5. Log energy

1. Line spectral frequencies (LSF)
2. Differential LSF, the successive differences of LSF
3. LSF with the zero crossing count of the filtered input signal
4. LSF with Linear prediction zero crossing ratio, the ratio of the zero crossing count (ZCC) of the input and the ZCC of the output of the LP analysis filter

—

Environmental sound recognition: a survey sachin chachada

1. Zero crossings
2. Amplitude
3. Power
4. Auto-regression
5. Adaptive time frequency decomposition
6. Short time Fourier
7. Brightness
8. Tonality
9. Loudness
10. Pitch
11. Chroma
12. Harmonicity
13. Perceptual filter bank
14. Advanced auditory model
15. (Cepstral) auto-regression
16. Rythm
17. Phase space

18. Eigen domain

—

6 Τι άλλο...;

Αναφορές

[1] Wenlei Shi and Xinhai Fan. Speech classification based on cuckoo algorithm and support vector machines. *2nd IEEE International Conference on Computational Intelligence and Applications*, 2017.