Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

«**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО**»

(Университет ИТМО)

Факультет программной инженерии и компьютерной техники

Пояснительная записка к курсовому проекту

По предмету «**Нейротехнологии и аффективные вычисления**»

Тема проекта: «**Сравнение эффективности различных классификаторов при распознавании эмоций по речи**»

Выполнили:

Ян Цзяфэн гр. P33212

Нгуен Чан Минь гр. P33211

Проверил:

Лямин Андрей Владимирович

Санкт-Петербург, 2020

Содержание

Команда проекта

Ян Цзяфэн: Использовать метод openSmile для извлечения наборов признаков.
Нгуен Чан Минь: Использовать метод Librosa для извлечения наборов признаков.

Аннотация

В настоящее время развитие распознавания речи можно считать очень зрелым, но это пока далеко от нашей цели - естественного взаимодействия человека и компьютера. Одна из причин этого заключается в том, что машины все еще не могут понимать наши эмоции, когда мы говорим. И это важная мотивация для исследования распознавания речевых эмоций. В данном проекте в основном сравниваются эффективности различных классификаторов при распознавании эмоций по речи.

Введение

Распознавание речи всегда было ключевой областью исследований искусственного интеллекта, а также одним из основных направлений будущих промышленных приложений технологий искусственного интеллекта. Распознавание речи не только открывает новый и преобразующий способ взаимодействия человека с компьютером, но также способствует эмоциональному общению между людьми и машинами. Распознавание эмоций по речи меняет наше общение с пользователями.

Распознавание эмоций по речи обычно относится к процессу, с помощью которого машины автоматически распознают человеческие эмоции и связанные с ними состояния по речи. Этот процесс включает три важных аспекта: правильную подготовку базы данных для оценки производительности системы; выбор подходящих признак для представления речи; использование соответствующих классификаторов.

Обычно, распознавание эмоций речи в основном извлекает следующие признаки: просодические (Pitch, Energy, Duration), спектральные (MFCC, LPCC, LFPC, GFCC, Formants), признаки качества голоса(Jitter, Shimmer, HNR, Normalized Amplitude Quotient, Quasi Open Quotient), признаки на основе Teager Energy Operator(TEO-FM-Var, TEO-Auto-Env, TEO-CB-Auto-Env). С момента развития распознавания речевых эмоций извлечение признаков все еще находится на стадии развития. Самые популярные наборы признаков - это набор признаков, выбранные организатором в INTERSPEECH 2009 Emotion Challenge и INTERSPEECH 2013 Paralinguistics Challenge, а также набор признаков в инструменте обработки речевых сигналов openSMILE.

Этап построения классификатора - это часть машинного обучения. В прошлом широко использовались классические методы машинного обучения, такие как модель гауссовой смеси (GMM), скрытая марковская модель (HMM) и машина опорных векторов (SVM). Благодаря развитию нейронных сетей такие методы, как модели долгосрочной и краткосрочной памяти (LSTM) и механизмы внимания, заменили классические методы и стали широко распространенными. В последние годы начали применяться сквозные методы, упрощающие или даже исключающие этапы извлечения признаков.

В этом проекте в основном сравниваются эффективности моделей, полученных при

обучении классификаторов SVM, MLP и CNN с помощью наборов признаков, извлеченных openSmile и Librosa.

Цель

Исследование преимуществ различных классификаторов на основе разных наборов данных при распознавании эмоций по речи.

Задачи

1. Собрать классические наборы данных, содержащих эмоциональную речь.
2. Обработка и анализ аудиофайлов для извлечения признаки.
3. Реализовать различные классификаторы временных рядов и обучить модели.
4. Тестировать модели.
5. Сравнить и анализировать эффективность каждого классификатора.

База данных

RAVDESS(The Ryerson Audio-Visual Database of Emotional Speech and Song)

Английский, около 1500 аудиозаписей от 24 человек (12 мужчин и 12 женщин), включая 8 различных эмоций (третье число в имени файла представляет эмоциональный тип): 01 = нейтральный, 02 = спокойный, 03 = счастливый, 04 = грустный, 05 = злой, 06 = испуганный, 07 = отвращение, 08 = удивленный.

Главная идея

Базовый процесс распознавания эмоций по речи показан на следующем рисунке:

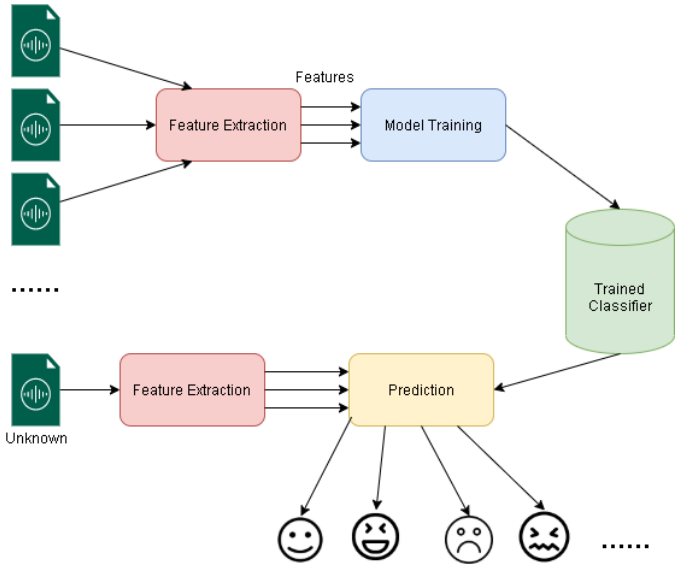

рис. 1 Блок-схема

Речевой сигнал сначала преобразуется в различные читаемые физические характеристики (высоту тона, энергию и т.д.) с помощью системы обработки речи. Каждый сегмент речевого

сигнала имеет свои уникальные характеристики. Некоторые из этих характеристик будут искусственно выбраны, извлечены системой, введены в предварительно обученный классификатор для различения и выведены в результате эмоционального состояния.

Извлечение признаков

На этапе извлечения признаков наиболее часто используемыми характеристиками речи являются энергия, высота тона и кепстральный коэффициент частоты Mel (MFCC). С момента развития распознавания речевых эмоций, извлечение признаков все еще находится на стадии развития.В настоящее время наиболее популярными методами извлечения признаков являются:

- использовать библиотеку обработки звука Librosa для извлечения некоторых часто используемых акустических характеристик, включая просодические характеристики (продолжительность голоса, длительность светового тона, эффективная продолжительность речи, кратковременная энергия, частота пересечения нуля, основная частота, функции LLD на основе кадров, такие как логарифмическая энергия и уровень звукового давления, и функции HSF, основанные на глобальных статистических значениях этих LLD для всей речи, включая Минимум, максимум, диапазон, среднее значение, стандартное отклонение, асимметрия и эксцесс, а также одна и та же операция для первой и второй разностей), функции спектральной корреляции (39-мерные функции MFCC) и характеристики качества звука (Центральная частота форманты 1/2/3 и ее полоса пропускания, возмущение частоты, возмущение амплитуды).
  feature = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
  feature=np.mean(librosa.feature.chroma_stft(S=np.abs(librosa.stft(X)),sr=sample_rate).T,axis=0)
  feature = np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T, axis=0)

- прямой вызов файлов конфигурации функции в openSMILE, включая набор признаков 2016-eGeMAPS(всего 88 признаков), набор признаков 2016-ComParE(всего 6373 признаков), 2009 г. -InterSpeech Emotion Challenge набор признаков (IS09_emotion)(всего 384 признаков). openSMILE - это программное обеспечение с доступным исходным кодом для автоматического извлечения признаков из аудиосигналов и для классификации речевых и музыкальных сигналов. openSMILE в основном применяется в области автоматического распознавания эмоций и широко используется в сообществе исследователей аффективных вычислений.
  Команда cmd: *SMILExtract_Release -C config_file -I input_file -O output_file*
  где, config_file: путь к выбранному файлу набора признаков opensmile,
  input_file: путь к входному аудиофайлу,
  output_file: путь к файлу результата вывода извлечения признаков

Алгоритм классификации

В этом проекте для обучения и получения моделей используются классификаторы SVM, MLP и CNN.

- Машины опорных векторов (SVM) - это модель с двумя классификациями.Его базовая модель - линейный классификатор с наибольшим интервалом, определенным в пространстве признаков. Реализуем svm-классификатор на основе sklearn.svm.SVC:
  class sklearn.svm.**SVC**(*, *C=1.0*, *kernel='rbf'*, *degree=3*, *gamma='scale'*, *coef0=0.0*, *shrinking=True*, *probability=False*, *tol=0.001*, *cache_size=200*, *class_weight=None*, *verbose=False*, *max_iter=-1*, *decision_function_shape='ovr'*, *break_ties=False*, *random_state=None*)

  На следующем рисунке показана точность модели, полученной путем обучения классификатора SVM с помощью набора признаков, извлеченных openSmile. (Стандарт сравнения: точность модели на обучающем наборе и точность модели на тестовом наборе)

```
from sklearn.svm import SVC
import numpy as np
X = np.asarray(X_train_IS09)
Y = np.asarray(Y_train_IS09)
clf_svc = SVC(decision_function_shape='ovo', kernel = 'rbf', C = 200,
          gamma = 0.00000001, probability = True, degree=0.1)
clf_svc.fit(X, Y)
print(clf_svc.score(X, Y))


from sklearn.metrics import accuracy_score
y_hat = clf_svc.predict(X_test_IS09)
accuracy_score(y_hat, Y_test_IS09)
```

```
0.6507936507936508

0.4027777777777778
```

рис. 2 точность модели SVM(openSmile)

На следующем рисунке показана точность модели, полученной путем обучения классификатора SVM с помощью набора признаков, извлеченных librosa.

```
from sklearn import svm
clf = svm.SVC(C= 9, gamma = 0.001, decision_function_shape='ovo')
clf.fit(X_train, y_train)
print(clf.score(X_train, y_train))
print(clf.score(X_test, y_test))
```

```
0.9989626556016598
0.5798319327731093
```

рис. 3 точность модели SVM(librosa)

- MLP также известен как многослойный персептрон или искусственная нейронная сеть (ANN). Помимо входного и выходного уровней, он может иметь несколько скрытых слоев между ними. Если скрытого слоя нет, он может решить проблему линейно разделяемых данных.
class sklearn.neural_network.MLPClassifier(hidden_layer_sizes=(100,),activation='relu', solver='adam',alpha=0.0001,batch_size='auto',learning_rate='constant',learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True,random_state=None, tol=0.0001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_change=10)

На следующем рисунке показана точность модели, полученной путем обучения классификатора MLP с помощью набора признаков, извлеченных openSmile.

```
from sklearn.neural_network import MLPClassifier
mlp_classifier = MLPClassifier(alpha=0.912, batch_size=32, learning_rate='adaptive', max_iter=500)
mlp_classifier.fit(X, Y)
print(mlp_classifier.score(X, Y))
y_hat = mlp_classifier.predict(X_test_IS09)
accuracy_score(y_hat, Y_test_IS09)
```

```
0.7073412698412699

0.41203703703703703
```

рис. 4 точность модели MLP(openSmile)

На следующем рисунке показана точность модели, полученной путем обучения классификатора MLP с помощью набора признаков, извлеченных librosa.

```
from sklearn.neural_network import MLPClassifier
mlp_classifer = MLPClassifier(alpha=0.01, batch_size=32,
                              hidden_layer_sizes=(64,32,32,16,8),
                              learning_rate='adaptive', max_iter=200)
```

```
mlp_classifer.fit(X_train, y_train)
```

```
MLPClassifier(activation='relu', alpha=0.01, batch_size=32, beta_1=0.9,
    beta_2=0.999, early_stopping=False, epsilon=1e-08,
    hidden_layer_sizes=(64, 32, 32, 16, 8), learning_rate='adaptive',
    learning_rate_init=0.001, max_iter=200, momentum=0.9,
    n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5,
    random_state=None, shuffle=True, solver='adam', tol=0.0001,
    validation_fraction=0.1, verbose=False, warm_start=False)
```

```
from sklearn.metrics import accuracy_score
print(mlp_classifer.score(X_train, y_train))
predictions = mlp_classifer.predict(X_test)
print(accuracy_score(y_test, predictions))
```

```
0.5529045643153527
0.4474789915966387
```

рис. 5 точность модели MLP(librosa)

- Сверточная нейронная сеть (CNN) имитирует конструкцию биологического механизма визуального восприятия, который может выполнять обучение с учителем и обучение без учителя. Совместное использование параметров ядра свертки в скрытом слое и разреженность соединений между слоями позволяют сверточной нейронной сети выполнять функции топологии, подобные сетке, с небольшим объемом вычислений.
  Стандарт сравнения: функция потерь и точность модели. В машинном обучении разница между прогнозируемым значением отдельной выборки и истинным значением модели обычно называется потерями. Чем меньше потеря, тем лучше модель, а функция, используемая для вычисления потерь, называется функцией потерь.

```
import keras
import numpy as np
import matplotlib.pyplot as plt
import tensorflow as tf
from keras.preprocessing import sequence
from keras.models import Sequential
from keras.layers import Dense, Embedding, BatchNormalization
from keras.utils import to_categorical
from keras.layers import Input, Flatten, Dropout, Activation
from keras.layers import Conv1D, MaxPooling1D
from keras.models import Model
from keras.callbacks import ModelCheckpoint

model = Sequential()
for size in [5, 5]:
    model.add(Conv1D(filters = 32, kernel_size = size, padding='same',input_shape=(384, 1)))
    model.add(BatchNormalization(axis = -1))
    model.add(Activation('relu'))
    model.add(Dropout(0.2))
model.add(Flatten())
model.add(Dense(16))
model.add(BatchNormalization(axis = -1))
model.add(Activation('softmax'))
model.add(Dropout(0.2))
model.summary()
```

```
cnn_history = model.fit(X_train_IS09_cnn, Y_train_IS09, batch_size=32, epochs=50, validation_data=(X_test_IS09_cnn, Y_test_IS09))
```

На следующих рисунках показаны фукция потерь и точность модели, полученной путем обучения классификатора CNN с помощью набора признаков, извлеченных openSmile.

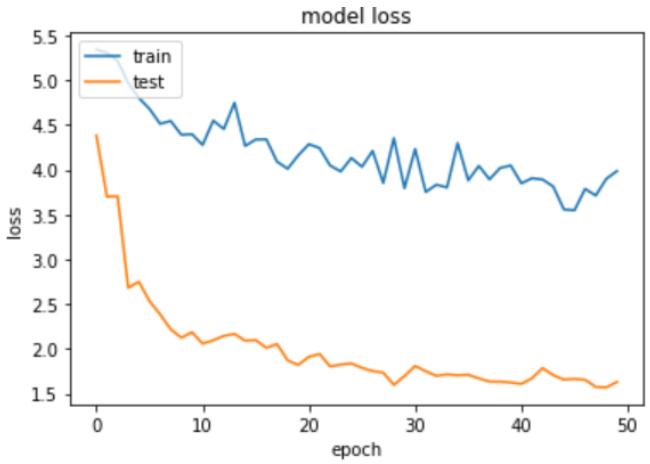

рис. 6 model loss CNN(openSmile)

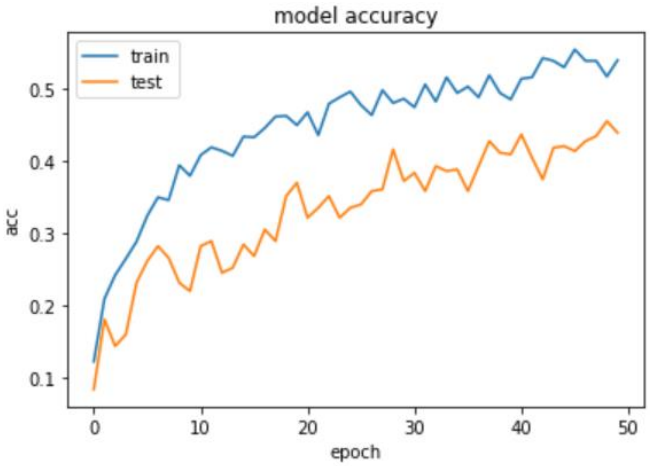

рис. 7 model accuracy CNN(openSmile)

На следующих рисунках показаны фукция потерь и точность модели, полученной путем обучения классификатора CNN с помощью набора признаков, извлеченных Librosa.

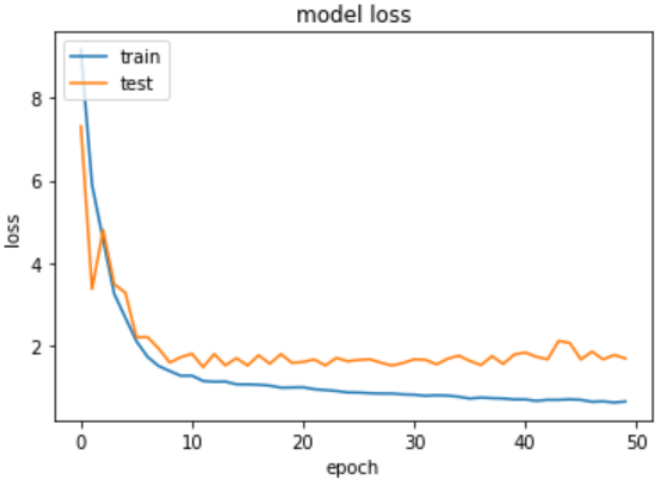

рис. 8 model loss CNN(librosa)

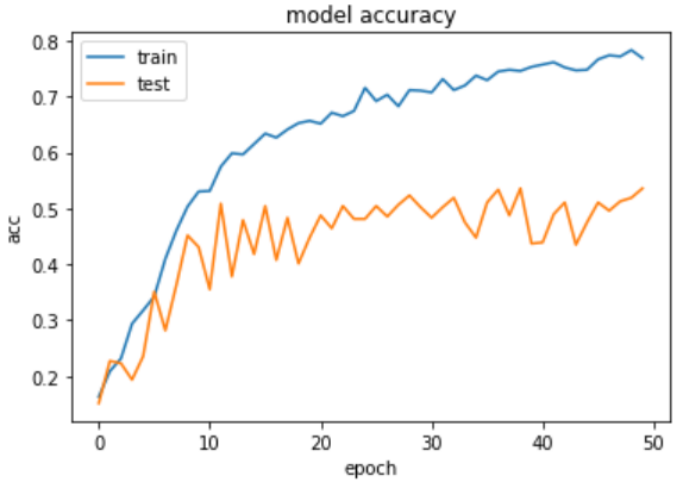

рис. 9 model accuracy CNN(librosa)

Результаты проекта

В следующей таблице показана точность каждого алгоритма классификатора на тестовом наборе для двух методов извлечения признаков.

	SVM	**MLP**	**CNN**
openSmile	0.40278	0.41204	0.4398
Librosa	0.57983	0.44748	0.5357

Выводы

В этом проекте openSmile может извлекать несколько типов наборов признаков для получения большего количества физических извлечений, представляющих аудиосигналы, но извлечение признаков занимает больше времени, а точность полученной модели ниже. Librosa извлекает один тип набора функций, время извлечения короче, а полученная модель более точна.

С точки зрения сравнения производительности классификаторов, точность модели CNN лучше, но время обучения модели больше.

Точность полученной модели все еще слишком низкая, поэтому в будущем для получения высокоточной модели необходимо постоянно корректировать параметры классификатора. Поскольку количество сравнительных примеров слишком мало, чтобы подтвердить вывод, в будущих исследованиях можно использовать несколько баз данных или несколько наборов признаков Librosa и openSmile.

Список литературы

1. Akçay, M. B., & Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, *116*, 56-76.
2. Website: https://github.com/Renovamen/Speech-Emotion-Recognition/blob/master/README_EN.md
3. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.