گزارش پروژه نهایی
درس شناسایی الگو

بهار ۹۹

فاطمه غفاری
۸۱۰۱۹۸۳۱۷

فهرست

مقدمه

صدای صحبت‌کردن انسان اطلاعات زیادی را در خود جای داده است. این اطلاعات می‌توانند در مورد صحبت بوده (زبان-شناختی) یا در مورد سخنران باشند (سن، جنسیت، ملیت و... .) تشخیص این اطلاعات برای انسان آسان است و این انگیزه‌ای برای ساخت ماشین‌هایی شده است که قادر به تشخیص این اطلاعات باشند.

کاربردهای زیادی را میتوان برای مسئله شناسایی صدای انسان متصور بود. مانند دیکته خودکار، کنترل‌های صوتی، احراز هویت و... . در این پروژه دو مسئله تشخیص جنسیت از روی صدا و خوشه بندی صداها بررسی شده اند.

مطالب فصل‌های آینده به این شرح اند: در فصل اول توصیفی آماری از داده‌هایی که در دست داریم ارائه می‌شود. در فصل دوم توضیحی درباره‌ی مراحل تمیزسازی داده‌ها و آماده‌سازی آن‌ها برای استخراج ویژگی داده می‌شود. در فصل سوم در مورد طبقه-بندی داده‌ها و بدست آوردن مدل، سپس تست کردن مدل و ارزیابی آن گفته می‌شود و درنهایت در فصل آخر، صداها خوشه بندی می‌شوند و جداسازی آن‌ها از یک دیگر بر اساس اندازه خوشه بررسی می‌شود.

بررسی آماری

در این بخش می‌خواهیم داده‌ها را بررسی کنیم تا ویژگی‌ها آماری آن را بدست آوریم.

به طور کلی ۲۸۵۹ فایل صوتی وجود دارد که مربوط به صدای ۲۹۲ نفر هستند. از هر فرد به طور میانگین حدود ۹,۷۹ فایل وجود دارد.

در شکل ۱ مشخص شده است که چند درصد فایل‌های صدا مربوط به خانمها و چند درصد مربوط به آقایان است.

```
Male percentage:  47.88387548093739 %

Female percentage:  52.11612451906261 %
```

شکل ۱: درصد خانمها و آقایان

مشاهده می‌شود که درصد صداهای مربوط به خانم‌ها و آقایان تقریبا برابر هستند. این ویژگی باعث می‌شود که داده‌های ما برای طبقه‌بند تشخیص جنسیت بسیار مناسب باشد.

شکل ۲ میانگین سنی افراد و بیشترین و کمترین سن را نشان می‌دهد.

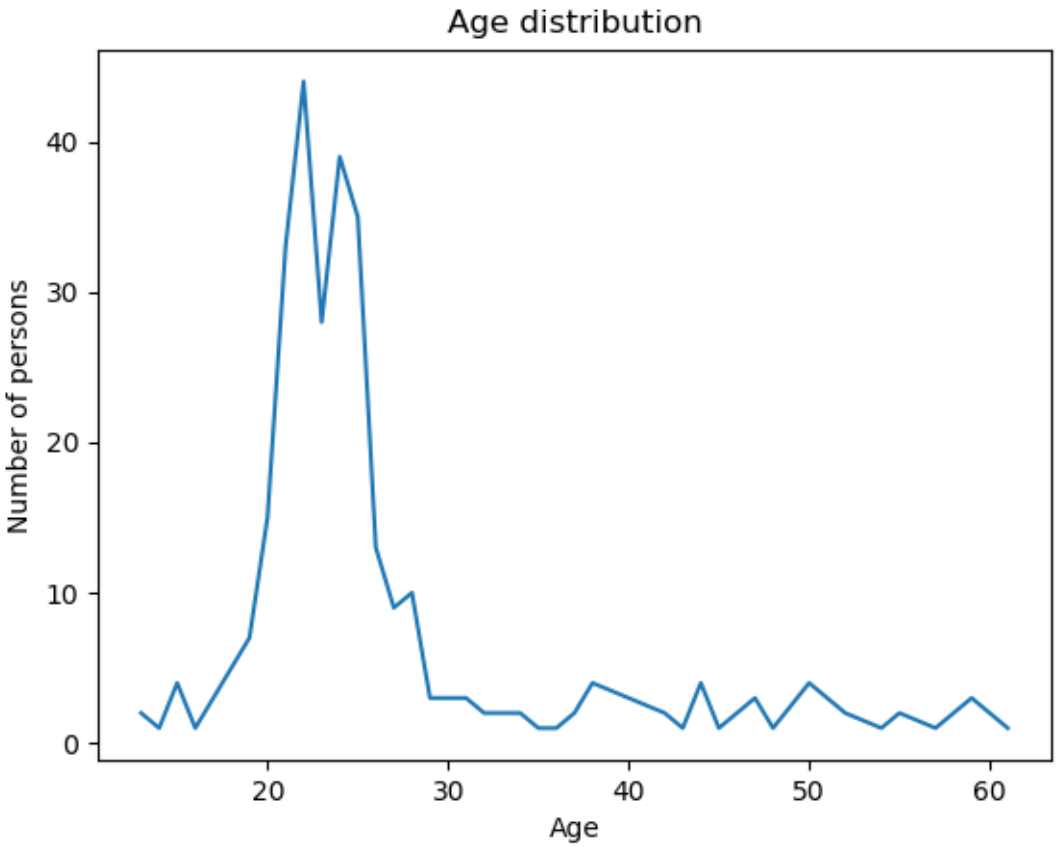
```
Average Age:  26.17808219178073

Min Age:  13

Max Age:  61
```

شکل ۲: میانگین و ماکسیمم و مینیمم سن

بازه سنی صاحبان صدا بازه بزرگی است اما میانگین آن بیشتر به مینیمم نزدیک است. به همین دلیل برای بررسی بیشتر نمودار شکل ۳ رسم شد.



شکل ۳: توزیع سنی صاحبان صدا

با توجه به نمودار مشخص می‌شود که سن اکثر صاحبان صدا در بازه ۲۰ تا ۳۰ است و تنها درصد کمی سن بیشتر یا کمتر از این بازه دارند. با توجه به این نتیجه، این داده‌ها برای بررسی‌های مربوط به سن مناسب به نظر نمی‌رسند. ازطرفی این نکته که در داده‌ها صدای کودک وجود ندارد یا درصد بسیار کمی وجود دارد نتیجه می‌شود که بررسی‌های مربوط به جنسیت احتمالا موفق خواهد بود.

تمیزسازی داده‌ها

دیتاست قرار داده شده در سایت شامل ۱۷ فولدر است که هر فولدر شامل تعدادی فولدر دیگر است که از ۱ شماره گذاری شده‌اند، به همراه یک فایل **CSV** که لیست فایل های موجود در آن فولدر و جنسیت و سال تولد هر گوینده در آن لیست شده است. از هر گوینده نیز یک یا چند فایل صوتی حدودا یک دقیقه‌ای وجود دارد.

ساختار ۱۷ دیتاستی که توسط افراد مختلف بارگذاری شده بودند تفاوت‌هایی با یکدیگر داشت که بدلیل محدود بودن تعداد آن‌ها این امکان وجود داشت که تغییراتی به صورت دستی در آن‌ها ایجاد شود که تاحدودی با یکدیگر همسان شوند.

ساختار فایل‌های **CSV** با یکدیگر متفاوت بود. در نتیجه تغییراتی حداقلی در آنها بوجود آوردم تا استفاده از آن‌ها در کد ساده باشد.

- هدر تمام فایل‌های **CSV** برابر با: **n, g, bd** برای شماره فایل، جنسیت و سال تولد قرار داده شد.
- در **CSV** شماره ۱۳ برای مرد از ۰ و برای زن از ۱ استفاده کرده بود و راهنمای این کد گذاری را نیز داخل خود **CSV** نوشته بود که این راهنمایی پاک شد و از کلمات **male** و **female** استفاده شد.
- خطوط اضافه آخر **CSV** ها حذف شد.
- فایل‌های دیتاست‌های شماره ۱، ۷، ۱۰ و ۱۵ به فرمت **xlsx** بودند که به **CSV** تبدیل شدند.
- در فایل شماره ۱۰ به جای سال تولد سن افراد نوشته شده بود که اصلاح شد.
- در برخی از فولدرها تعدادی فایل مانند **desktop.ini** و ازین قبیل وجود داشت که می‌توانست با فایل صدا اشتباه گرفته شود و منجر به خطا شود که حذف شد.
- در دیتاست شماره ۱۵، فولدر شماره ۸، فایل «کوچه مانی ۸ – ۲» خراب بود و حذف شد.

در ادامه باید فرمت و سمپل ریت تمام فایل‌ها یکسان سازی شود که هر دو این‌ها با دستور خواندن فایل صوتی توسط پکیج مربوطه انجام می‌شود. در بخش بعد در این مورد بیشتر توضیح داده خواهد شد.

سپس تلاش شد که لحظات سکوت و نویز از فایل‌های صوتی حذف شود اما با بررسی چندین فایل مشخص شد که تغییر چندانی حاصل نمی‌شود و فایل‌ها عموما بدون سکوت و نویز هستند. در نتیجه این بخش موقتا حذف شد تا در صورت نتیجه نامطلوب اعمال شود.

استخراج و انتخاب ویژگی‌ها

در جستجوهایی که در مرحله قبل انجام شد، تصمیم بر این شد که از ویژگی‌های **MFCC** برای این کاربرد استفاده شود که برای مصارف مرتبط با صدای انسان کاربرد زیادی دارد.

بدست آوردن ویژگی‌های **MFCC** به طور خلاصه به صورت زیر است:

- ابتدا صوت به بازه‌های زمانی کوتاه تقسیم می‌شود.
- از هر بازه طیف توان گرفته می‌شود.
- از جایی‌که طیف توان نمی‌تواند تفاوت بین فرکانس‌های نزدیک را بخوبی نشان دهد. در نتیجه برای درک اینکه واقعا چه مقدار انرژی در فرکانسی خاص وجود دارد، باید بین‌هایی از طیف توان با هم جمع شوند که این عمل به کمک **MEL filterbank** انجام می‌شود.
- لگاریتم انرژی‌های بدست آمده گرفته می‌شود. این کار دقیقا منطبق بر عملکرد گوش انسان است. یعنی انسان هم صداهایی با انرژی دو برابر را واقعا دو برابر بلندتر نمیـشنود.
- در نهایت از آن‌ها **DCT** گرفته می‌شود تا از هم مستقل شوند.

برای استخراج این ویژگی‌ها تعداد زیادی پکیج پایتون وجود دارد که برخی از آن‌ها عبارتند از: **librosa [1]**، **PyAudioAnalysis[2]** و ... . در جستجوهایی که در مورد این پکیج‌ها داشتم متوجه شدم که ظاهرا کاربران **librosa** تجربه بهتری داشته‌اند در نتیجه از همین پکیج برای استخراج ویژگی‌ها استفاده شده است.

از آن‌جایی که فرمت داده‌ها با یکدیگر متفاوت بود نیاز بود که ابتدا فرمت همه آن‌ها یکسان شود. اما با نصب **ffmpeg** که یک ابزار برای تبدیل فرمت صدا و ویدئو است و اضافه کردن آن به مسیر کامندلاین، **librosa** می‌تواند تمامی فرمت ها را بخواند و ویژگی‌های آن‌ها را استخراج کند و در نتیجه نیازی به تبدیل فرمت به صورت دستی نیست.

از آنجایی که تعداد فایل‌های صوتی نسبتا زیاد (۲۸۵۹ عدد) است، در صورتی که از سمپل ریت[1] بالایی استفاده شود، ویژگی‌ها قابل ذخیره در آرایه **numpy** نخواهند بود و در صورت استفاده از روش‌های دیگر برای ذخیره سازی نیز حجم بسیار بالایی اشغال

[1] **Sample rate**

خواهند کرد. در نتیجه در هنگام خواندن فایل‌های صوتی سمپل ریت آن‌ها برابر با ۸۰۰۰ هرتز در نظر گرفته شد که سمپل ریت پایین و در عین حال برای صحبت انسان قابل قبول است.

از آن‌جایی که اندازه ماتریس ویژگی‌های **MFCC** مستقیما به تعداد سمپل ها و در نتیجه طول فایل صدا بستگی دارد همه فایل‌ها باید طول یکسانی داشته باشند. تمام فایل‌ها طولی در حدود یک دقیقه داشتند که طول همه آنها یکسان و برابر ۶۰ ثانیه قرار داده شد. تابع مربوط به استخرج ویژگی‌های **MFCC** به طور پیش‌فرض ۲۰ بردار **MFCC** را برمی‌گرداند. از این ۲۰ بردار در کاربردهای دیکته خودکار و شبیه به آن استفاده می‌شود، در نتیجه می‌توان نتیجه گرفت که این ۲۰ بردار مربوط به ویژگی‌های زبانشناسی هستند که از ویژگی‌های شخصی گویند مانند سن و جنس مستقل است. در نتیجه برای این کاربرد از ۴۰ ویژگی اول **MFCC** استفاده می‌کنیم.

در نهایت ویژگی‌های **MFCC** برای تمامی فایل‌های صدا استخراج شد و به همراه برچسب‌های سال تولد و جنسیت و هویت برای استفاده های بعدی ذخیره شد.

طبقه بندی

در بررسی‌های مرحله‌های قبل به این نتیجه رسیدم که طبقه‌بندهایی که معمولا با ویژگی‌های **MFCC** استفاده می‌شوند و بهترین عملکرد را دارند **GMM** و **HMM** هستند. در نتیجه در این مرحله از طبقه بند **GMM** استفاده شد. برای ارزیابی عملکرد طبقه بند از **10-fold cross validation** استفاده شد. یعنی کل داده ها به ده قسمت تقسیم شد و در در هر مرحله یک قسمت به عنوان داده تست و نه قسمت دیگر به عنوان داده‌های ترین مورد استفاده قرار گرفتند. تعداد کامپوننت های گوسی دو عدد در نظر گرفته شد. به جهت بزرگ بودن ویژگی ها برای برنخوردن به خطای حافظه، **covariance_type** به صورت **diag** استفاده شد. یعنی فرض شده که ماتریس کوواریانس هر کدام از کامپوننت‌های گوسی قطری هستند که باعث سادگی مسئله می‌شود. البته ممکن است در عملکرد سیستم ایجاد مشکل کند که در ادامه مشخص می‌شود که این اتفاق نمی‌افتد.

شکل‌های زیر دقت و ماتریس **confusion** میانگین **10-fold** را نشان می‌دهند.

```
10-fold average accuracy:  94.67390504232611
10-fold average confusion matrix:
[[1307.   62.]
 [  90. 1400.]]
```

*شکل ۴: دقت و ماتریس **confusion***

مشاهده می‌شود که دقت بدست آمده ۹۴,۶۷ درصد است که نتیجه بسیار خوبی به حساب می‌آید. با توجه به مناسب بودن داده‌ها از نظر آماری و همچنین مناسب بودن ویژگی‌ها و طبقه‌بند برای این کاربرد، از ابتدا می‌شد حدس زد که در این بخش نتیجه خوبی حاصل شود.

خوشه بندی

برای خوشه بندی از دو روش **k-means** و **gmm** استفاده شد اما در تعداد خوشه‌های بالا تغییرات چشمگیری بین آنها مشاهده نشد.

برای بررسی عملکرد خوشه‌بندی، از معیارهای متفاوتی استفاده شد که در ادامه توضیح داده می‌شود.

- **Adjusted rand index**: تابعی است که شباهت بین برچسب‌های حقیقی داده‌ها و برچسب‌های خوشه بندی آنها را مقایسه می‌کند و عددی بین ۰ تا ۱ بازمی‌گرداند که ۱ نشان‌دهنده بیشترین شباهت است. تغییر لیبل‌ها تاثیری در این امتیاز ندارد.

$$RI = \frac{a+b}{C_2^{n_{sample}}} \quad , \quad ARI = \frac{RI - E[RI]}{max(RI) - E[RI]}$$

C: برچسب‌های حقیقی داده‌ها

a: تعداد سمپل‌هایی که واقعا در یک کلاس هستند و در یک خوشه هستند.

b: تعداد سمپل‌هایی که در دو کلاس و همچنین دو خوشه متفاوت هستند.

- **Adjusted mutual information**: این تابع بر اساس اطلاعات مشترک بین برچسب خوشه بندی و برچسب-های حقیقی است.

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} P(i,j)log\left(\frac{P(i,j)}{P(i)P'(j)}\right) \quad , AMI$$

$$= \frac{MI - E[MI]}{mean\big(H(U), H(V)\big) - E[MI]}$$

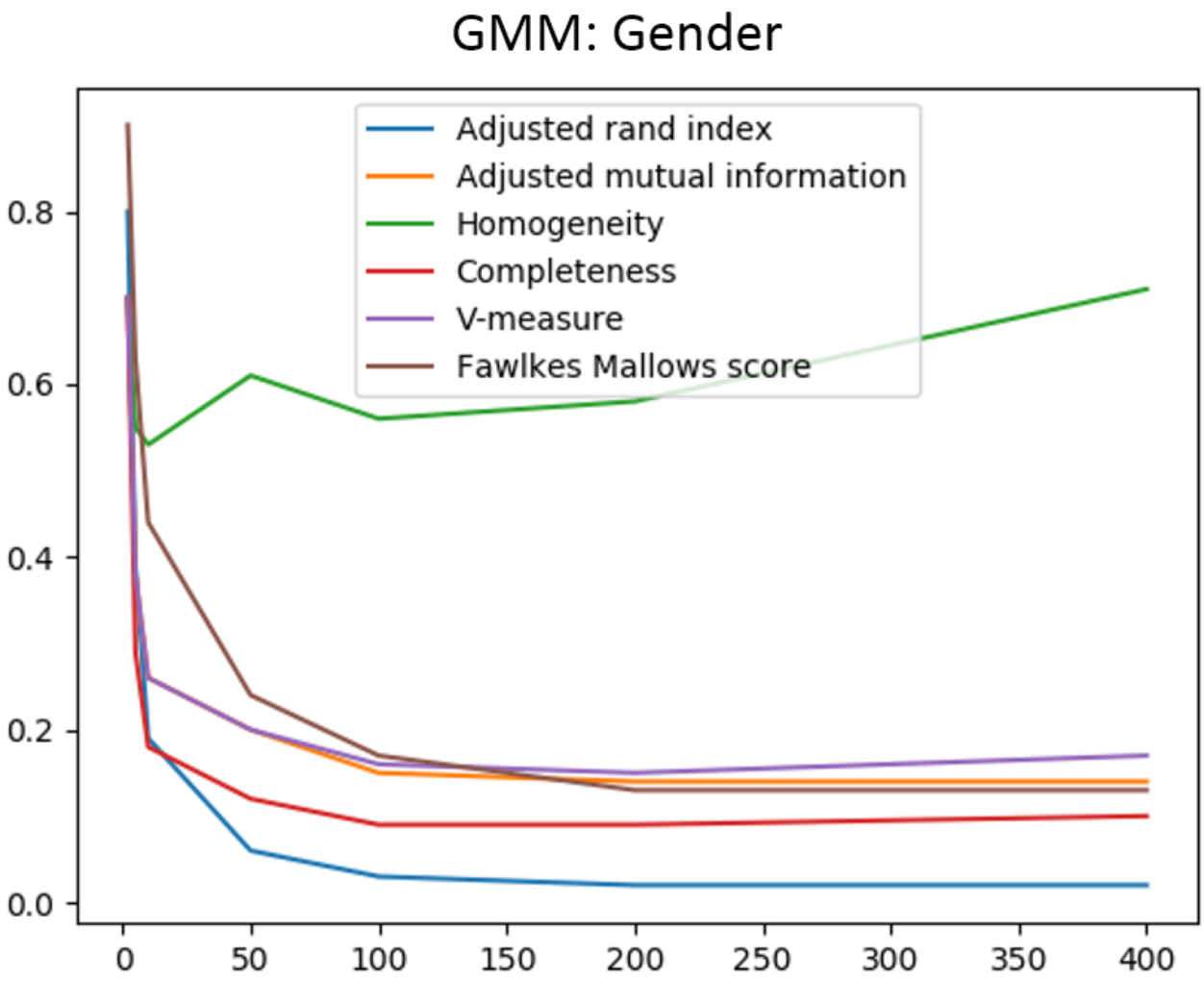
- **Homogeneity**: اگر هر کلاستر تنها شامل اعضای یک کلاس باشد این مقدار ۱ می‌شود.

$$h = 1 - \frac{H(C|K)}{H(C)}$$

- **Completeness**: اگر تمام اعضای هر کلاس عضو یک کلاستر باشند این معیار ۱ می شود.

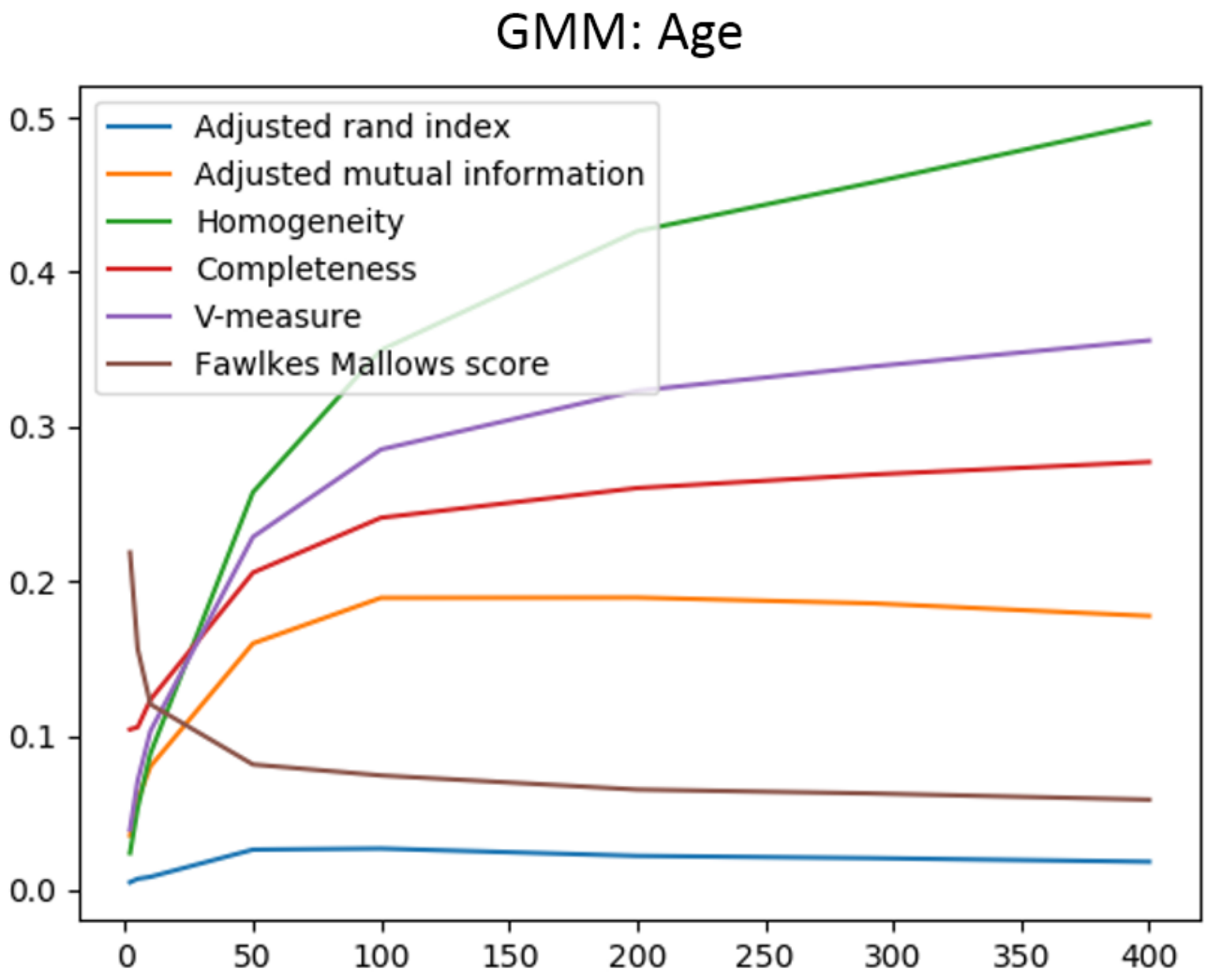
$$c = 1 - \frac{H(K|C)}{H(C)}$$

حال با توجه به معیارهای بالا، نمودار جنسیت، سن و هویت نسبت به تعداد خوشه ها در زیر رسم شده است:



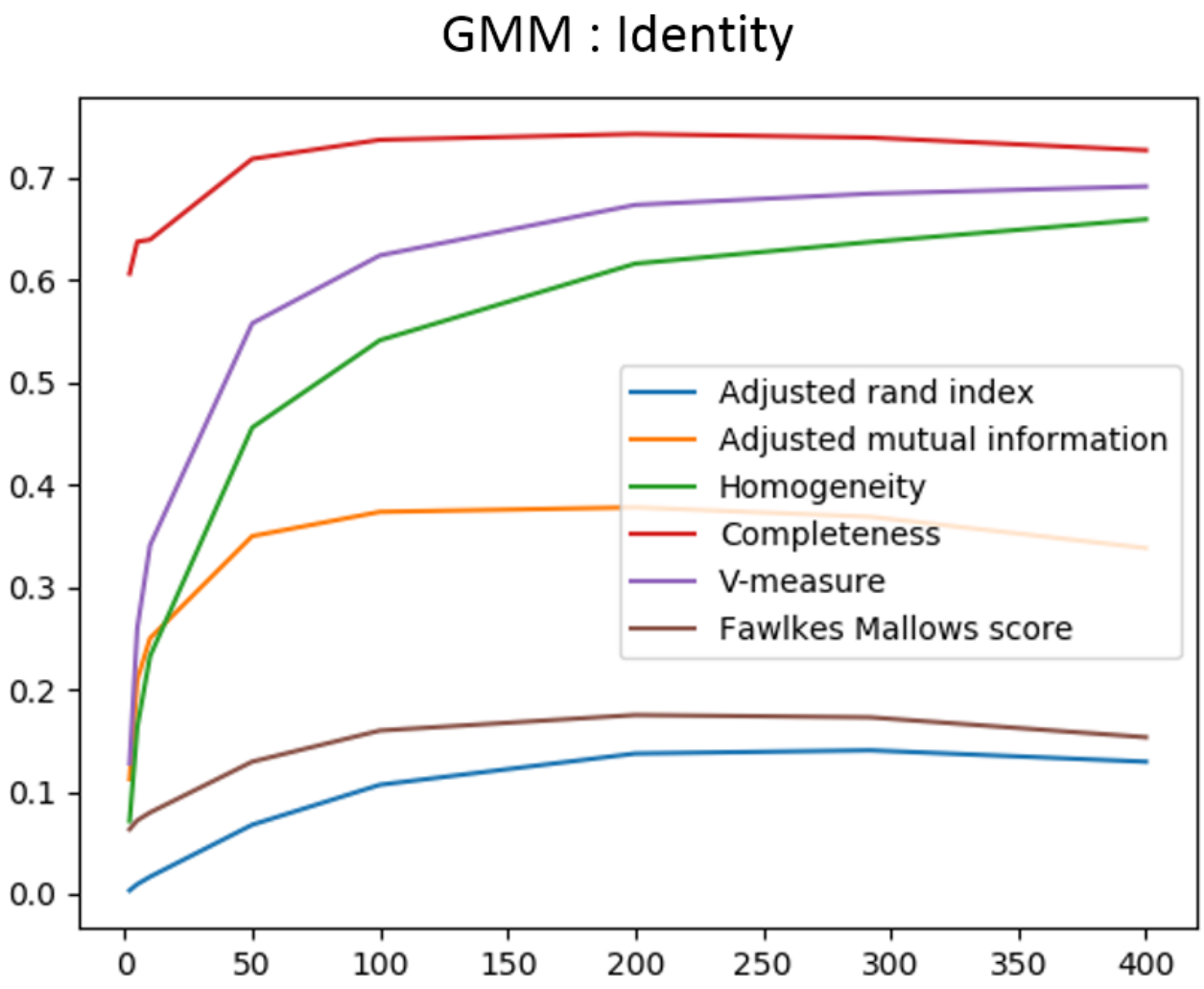
شکل ۵: مقدار معیارها به نسبت تعداد خوشه‌ها برای جنسیت با ***GMM***

با توجه به نمودار بالا مشاهده می‌شود که در کمترین مقدار که ۲ خوشه است سیستم عملکرد ایده‌آل داشته است و کاملا جنسیت را جدا کرده است اما با بیشتر شدن تعداد خوشه ها این عملکرد افت کرده است. که همان نتیجه ای است که انتظار آن را داشتیم.



شکل ۶: مقدار معیارها به نسبت تعداد خوشه‌ها برای سن با ***GMM***

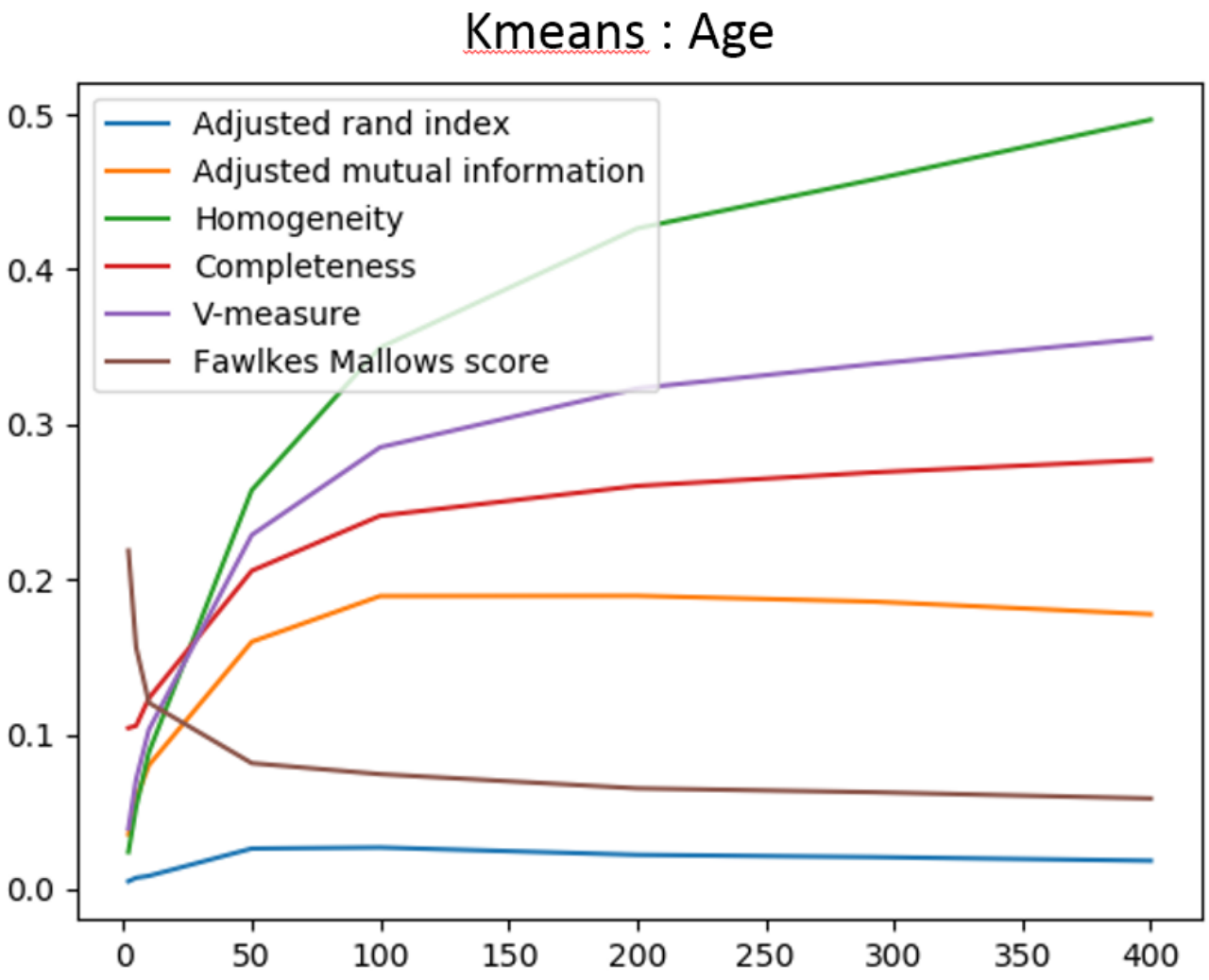
مشاهده می‌شود که با زیاد شدن تعداد خوشه ها عملکرد سیستم در جداسازی سنی بهبود می‌یابد اما از آن جایی که با زیاد شدن خوشه ها تعداد سمپل در هر خوشه کاهش می‌یابد که خواهیم دید که بیشتر آنـها متعلق به یک نفر هستند طبیعی است که عملا سمپلـها از نظر سنی هم جدا شوند. در بخش تحلیل آماری داده‌ها هم دیده بودیم که نمیتوانیم انتظار عملکرد خوبی برای تشخیص سن داشته باشیم.



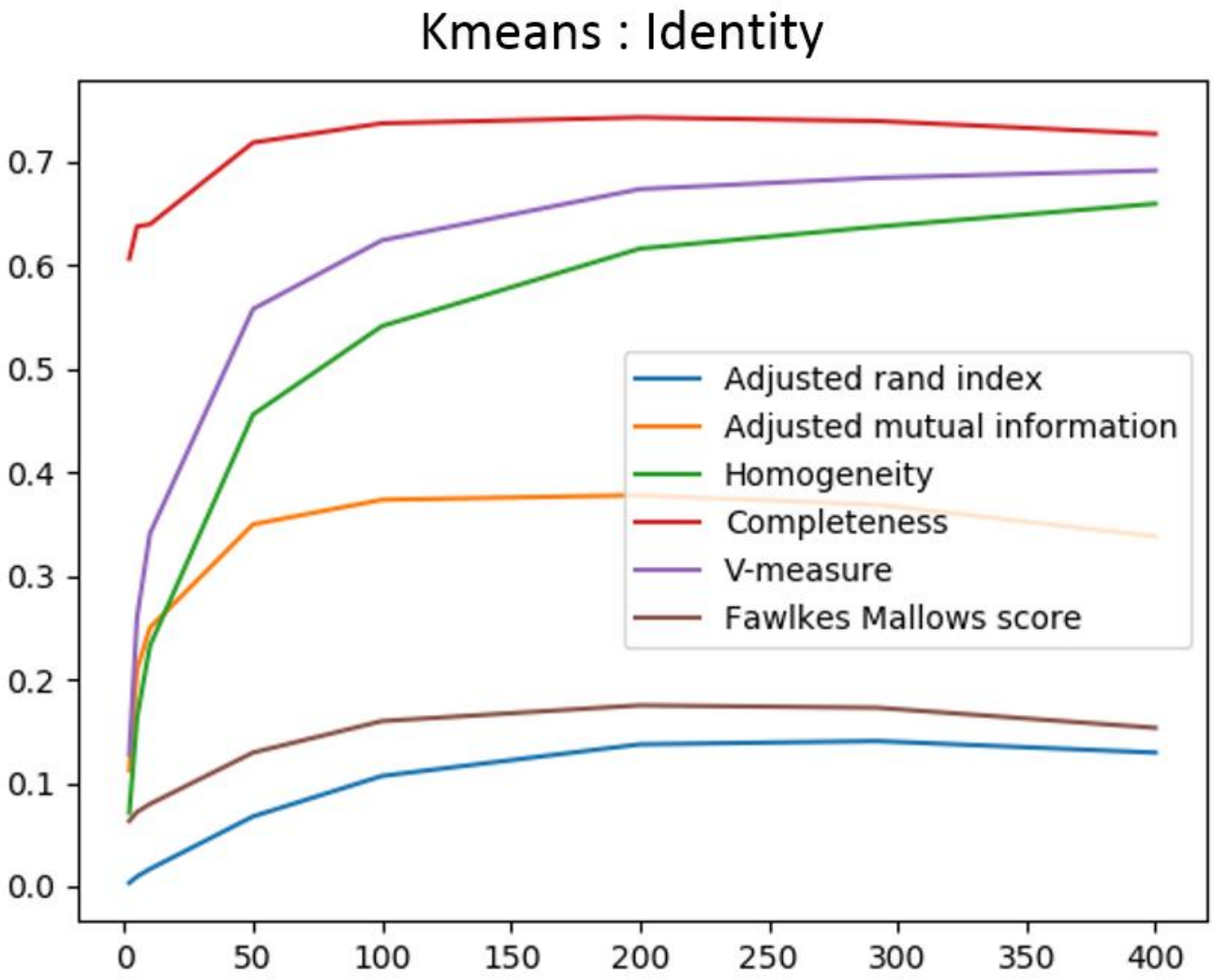
شکل ۷: مقدار معیارها به نسبت تعداد خوشه‌ها برای هویت با ***GMM***

در رابطه با هویت مشاهده می‌شود که با این که معیارهایی که بر پایه شباهت دو سری برچسب حقیقی و پیش بینی شده‌اند مقادیر بالایی ندارند اما در تعداد خوشه‌های بالا معیارهای **completeness** و **homogeneity** مقادیر قابل توجهی دارند. برای مثال در حدود ۳۰۰ خوشه که نزدیک به تعداد افراد (۲۹۸) است، می‌توان به صورت حدودی گفت که ۷۰ درصد از سمپلـهایی که متعلق به یک فرد است در یک خوشه هستند و حدود ۷۰ درصد از سمپل‌هایی که در یک خوشه هستند متعلق به یک نفرند.

در ادامه نتایج مربوط به خوشه بندی **Kmeans** آمده است که تفاوت چندانی با نتایج خوشه بندی **GMM** ندارند.



شکل ۸: مقدار معیارها به نسبت تعداد خوشه‌ها برای سن با Kmeans



شکل ۸: مقدار معیارها به نسبت تعداد خوشه‌ها برای هویت با **Kmeans**

منابع

[1] McFee, Brian, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg and O. Nieto, "librosa: Audio and music signal analysis in python," in *python in science conference*, 2015.

[2] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," *PloS one*, vol. 10, no. 12, 2015.