研究雑まとめ
MBTI データセットにおける特徴分析

田中直哉

1　研究背景および目的

近年 SNS 上で 16 性格診断と題した高精度な性格診断として MBTI が存在する。

しかし Kaggle 上に存在する MBTI のデータセットは本来の質問数である 93 個の特徴量に対して 8 個しか存在しない。

1.1　データセットについて

Kaggle より引用(Google 翻訳を使用)

説明

この合成データセットは、人口統計学的要因、興味分野、および性格スコアの組み合わせに基づいて、マイヤーズ・ブリッグス・タイプ指標（MBTI）の性格タイプを探索および予測するために設計されています。10 万以上のサンプルが含まれており、それぞれが MBTI タイプを決定する様々な特徴を持つ個人を表しています。このデータセットは、異なる性格特性と、年齢、性別、教育、興味などの外的要因との相関関係を研究するために使用できます。

データセットの説明

Age:個人の年齢を表す連続変数。

Gender:個人の性別を示すカテゴリ変数。可能な値は「男性」と「女性」です。

Education:バイナリ変数。値が 1 の場合、個人は少なくとも大学院レベル（またはそれ以上）の教育を受けていることを示し、値が 0 の場合、大学、高校レベル、または無学であることを示します。

Interest:個人の主な興味分野を表すカテゴリ変数。

Introversion Score： 0 から 10 までの連続変数で、個人の内向性または外向性の傾向を表します。スコアが高いほど、外向性の傾向が強いことを示します。

Sensing Score： 0 から 10 までの連続変数で、感覚と直観のどちらを重視するかを表します。スコアが高いほど、感覚が重視されることを示します。

Thinking Score： 0 から 10 までの連続変数で、思考と感情のどちらを優先するかを示します。スコアが高いほど、思考を優先する傾向があります。

Judging Score： 0 から 10 までの連続変数で、知覚よりも判断を好む傾向を表します。スコアが高いほど、判断を好む傾向を示します。

Personality:人物パーソナリティタイプを含むターゲット

引用終わり

表 1-1　データセット最初の 5 行

Age	Gender	Education	Introversion Score	Sensing Score	Thinking Score	Judging Score	Interest	Personality
21.0	Female	1	5.89208	2.144395	7.32363	5.462224	Arts	ENTP
24.0	Female	1	2.48366	3.206188	8.06876	3.765012	Unknown	INTP
26.0	Female	1	7.02910	6.469302	4.16472	5.454442	Others	ESFP
30.0	Male	0	5.46525	4.179244	2.82487	5.080477	Sports	ENFJ
31.0	Female	0	3.59804	6.189259	5.31347	3.677984	Others	ISFP

1.2 MBTI の特徴量

先述した通り MBTI は 93 個の質問によって精度が 100%となっているが本データセットは 8 個の特徴量であらわされているため、その特徴量の妥当性を検証する。また、質問と関係ない特徴量に対して MBTI への影響度を測定することに加えデータ構造を分析することで各変数の特徴を分析する。

2 研究手法

2.1 精度の測定

クロスバリデーションで 500 分割して精度の検証を行う。

目的としては特徴量の妥当性および 2.2 項における分類寄与率の測定において精度を担保する上で測定を行う。

2.2 分類寄与率の測定

GBDT は決定木系アルゴリズムであるため一つの変数でどれだけ分類できるか測定することができる。

そこで、各変数が MBTI を分類するうえでの影響度を測定する。

2.3 間違った傾向についての分析

Hold-out 方式で訓練データ 70%でテストデータ 30%とし、2.1 項において精度の代表値に近い精度を出したときに Classification_Report(各目的変数の値毎に)と混合行列を出力し、誤った傾向について考察する。

2.4 MBTI 別特徴量のヒストグラム

MBTI 別に特徴量をヒストグラムで画像を保存する。この画像を MBTI 別に比較することで MBTI 別の特徴量の傾向を分析する。

3　結果

3.1　精度の測定

クロスバリデーションで 500 回精度の検証を行った結果得られたようやく統計量を表 3-1 に記す。また、ヒストグラムを図 3-1 に示す。

表 3-1　500 回クロスバリデーションした結果得られたようやく統計量

	cross_validation
count	500.000000
mean	0.895525
std	0.032741
min	0.806818
25%	0.873563
50%	0.896552
75%	0.919540
max	0.988506

表 3-1 の結果から平均値および中央値ともに約 89%という事が分かる。

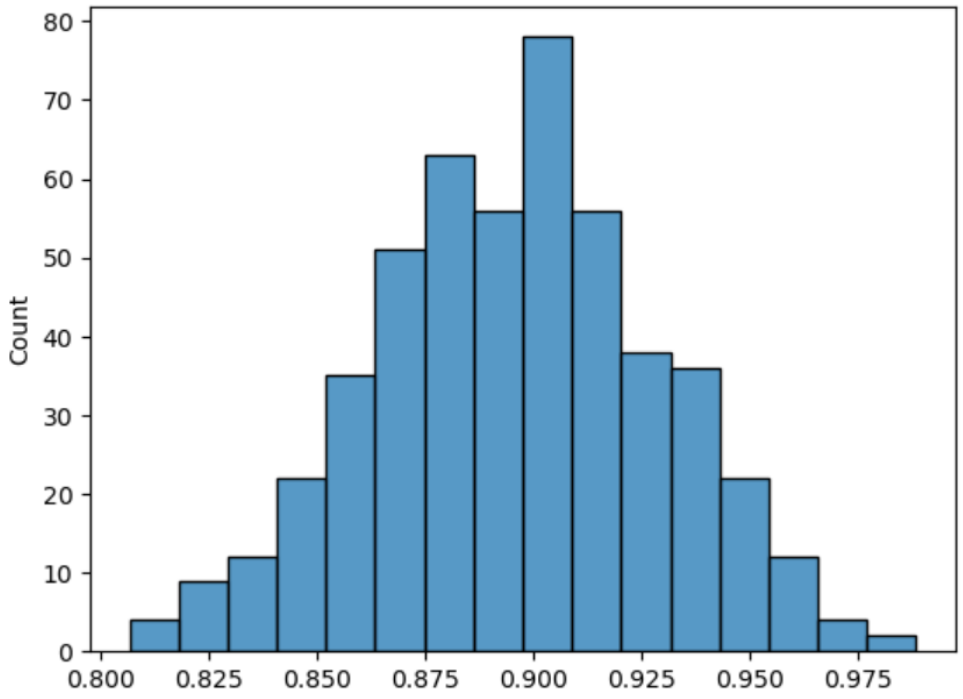

図 3-1　500 回クロスバリデーションを行った際に得られた精度のヒストグラム

図 3-1 の結果から 500 回クロスバリデーションを行い最頻値は 90%程度であると考えられる。

3.2　分類寄与率の測定

GBDT で MBTI における各特徴量について 89%の信頼度で分類寄与率を測定した結果を表 3-2 に記す。

表 3-2　GBDT を用いた分類寄与率(信頼度 90%)

	imp
Thinking Score	0.199917
Judging Score	0.196146
Introversion Score	0.186438
Sensing Score	0.175604
Age	0.105313
Gender_Male	0.034042
Education	0.032250
Interest_Arts	0.023229
Interest_Sports	0.020646
Interest_Unknown	0.011729
nterest_Technology	0.008146
Interest_Others	0.006542

この結果から各種スコアと年齢が MBTI に関係していることが考えられる。

一方で教育や性別、興味などは MBTI の分類結果に寄与しにくいことも考えられる。

3.3　間違った傾向についての分析

精度が代表値だった時の Classification_Report を図 3-2 に、混合行列を表 3-3 に記す。

	precision	recall	f1-score	support
ISFP	0.83	0.91	0.87	278
ISTP	0.89	0.94	0.91	277
INTP	0.85	0.88	0.87	247
ENTP	0.92	0.85	0.88	284
INFJ	0.91	0.87	0.89	264
ESTP	0.90	0.90	0.90	272
ESFP	0.84	0.87	0.86	259
ENFJ	0.92	0.89	0.91	267
INFP	0.90	0.86	0.88	309
INTJ	0.92	0.87	0.89	290
ISTJ	0.85	0.97	0.91	261
ENFP	0.90	0.85	0.87	269
ESFJ	0.90	0.88	0.89	293
ISFJ	0.84	0.90	0.87	284
ENTJ	0.91	0.88	0.90	249
ESTJ	0.95	0.90	0.92	272
accuracy			0.89	4375
macro avg	0.89	0.89	0.89	4375
weighted avg	0.89	0.89	0.89	4375

図 3-2　MBTI を特徴量で分類予測した時の Classification_Report

f1 スコアに着目すると、どの MBTI も正解率の代表値程度(誤差±3 程度)に間違えていることが分かる。

表 3-3　MBTI を分類予測した際の混合行列

	ISFP	ISTP	INTP	ENTP	INFJ	ESTP	ESFP	ENFJ	INFP	INTJ	ISTJ	ENFP	ESFJ	ISFJ	ENTJ	ESTJ
ISFP	252	0	0	0	0	0	22	0	4	0	0	0	0	0	0	0
ISTP	1	259	1	0	0	15	0	0	0	0	1	0	0	0	0	0
INTP	0	12	217	15	0	0	0	0	2	1	0	0	0	0	0	0
ENTP	0	0	32	240	0	12	0	0	0	0	0	0	0	0	0	0
INFJ	0	0	0	0	230	0	0	18	1	1	0	0	0	14	0	0
ESTP	0	20	0	5	0	246	1	0	0	0	0	0	0	0	0	0
ESFP	31	0	0	0	0	0	225	0	0	0	0	3	0	0	0	0
ENFJ	0	0	0	0	16	0	0	238	0	1	0	0	9	2	1	0
INFP	17	0	1	0	0	0	1	0	267	0	0	23	0	0	0	0
INTJ	0	0	3	0	1	0	0	0	0	252	15	0	1	0	18	0
ISTJ	0	0	0	0	0	0	0	0	0	1	254	0	0	0	0	6
ENFP	0	0	0	2	0	0	17	0	22	0	0	228	0	0	0	0
ESFJ	0	0	0	0	0	0	1	1	0	0	0	0	258	33	0	0
ISFJ	1	0	0	0	5	0	0	1	0	0	1	0	20	256	0	0
ENTJ	0	0	0	0	0	0	0	0	0	19	2	0	0	0	220	8
ESTJ	0	0	0	0	0	0	0	0	0	0	25	0	0	0	2	245

間違えた傾向としては、まず I と E(内向的か外向的か)で間違えており、その次に N(直感型)と S(感覚型)で間違えている。

間違ている項目を見てみると I と E は多くて正解に対して 10%以上間違えており、S と N は間違えた場合は多くて 5%程度間違えている。

3.4　MBTI 別特徴量のヒストグラム

MBTI 別に各特徴量をヒストグラムにした結果を図 3-3 から図 3-x まで示す。

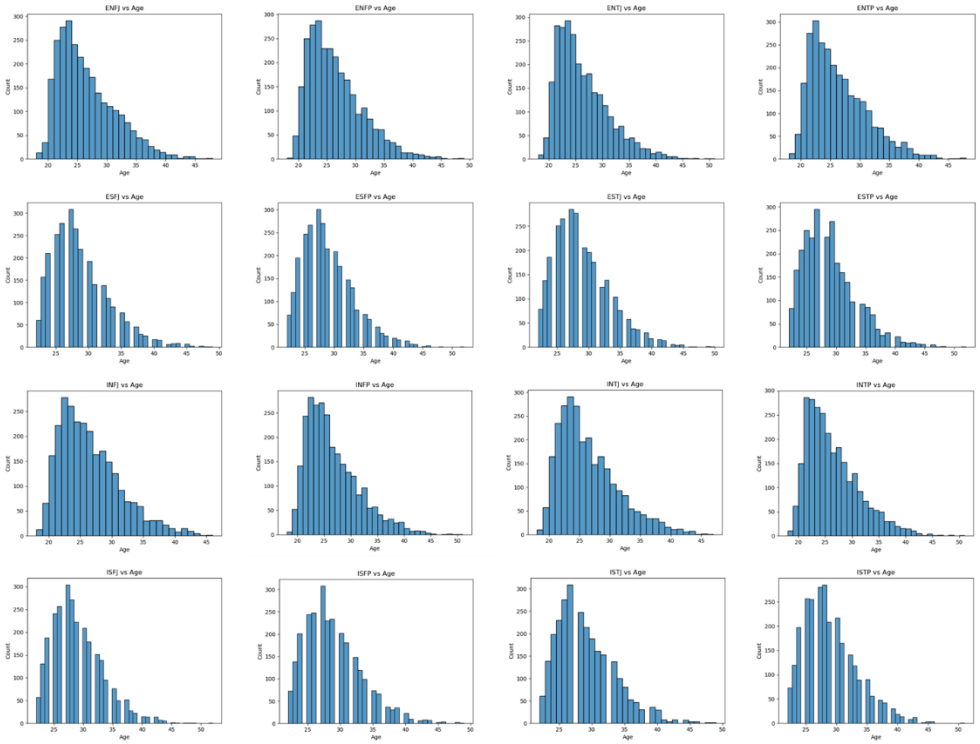

図 3-3　MBTI 別年齢

図 3-3 から MBTI において年齢による差は少ないが、分布から若年層を中心としたデータセットであることが分かる。

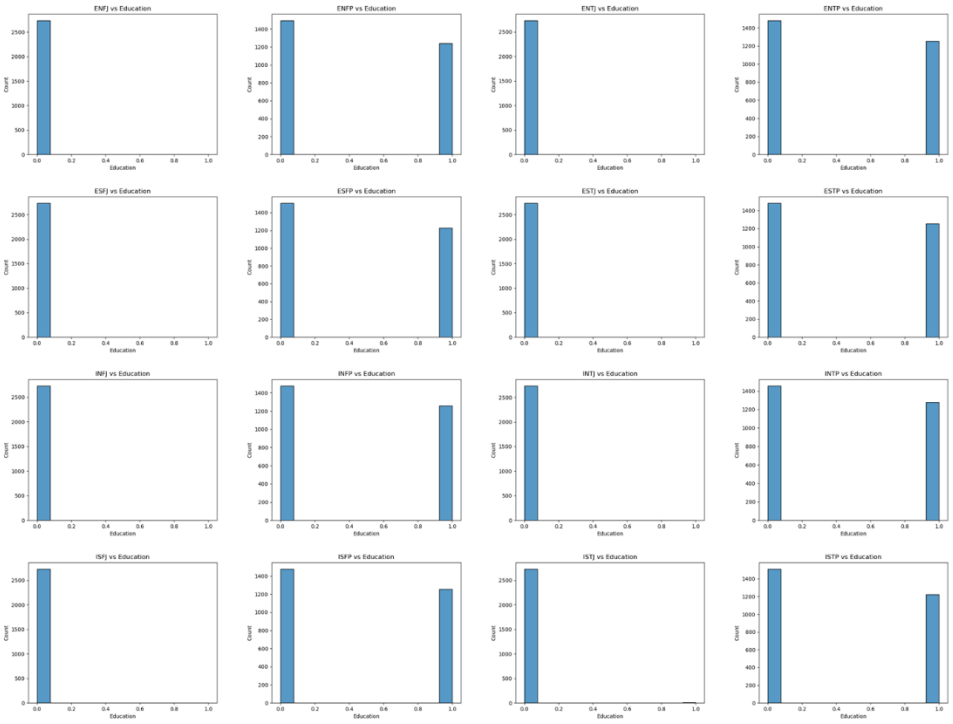

図 3-4　MBTI 別教育

本データセットにおいて J(判断型)と P(知覚型)では知覚型の P は大学院進学あるいはそれ以上の結果になり J は大学進学までという結果になった。

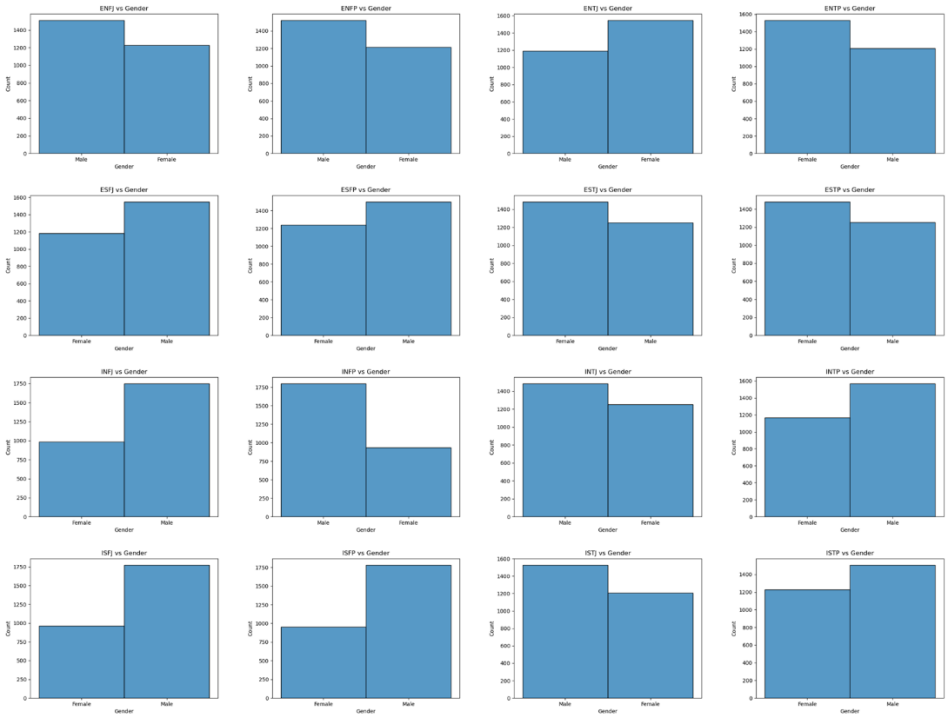

図 3-5　MBTI 別性別

E(外向的)の性別は S(感覚型)か N(直感的)かで男女比が逆転しており TJ(論理的思考)の場合、他の S か N と比率が逆転している。また、N の時は男性が多く、S の時は女性が多い。

また、I(内向的)の性別は INFJ と ISFJ で同じ性別の多さは同じ程度の比率だが E と比べて男女比が大きく異なる。同様に、INFP と ISFP は性別が逆転しているが E と比べて男女比が大きく異なる。

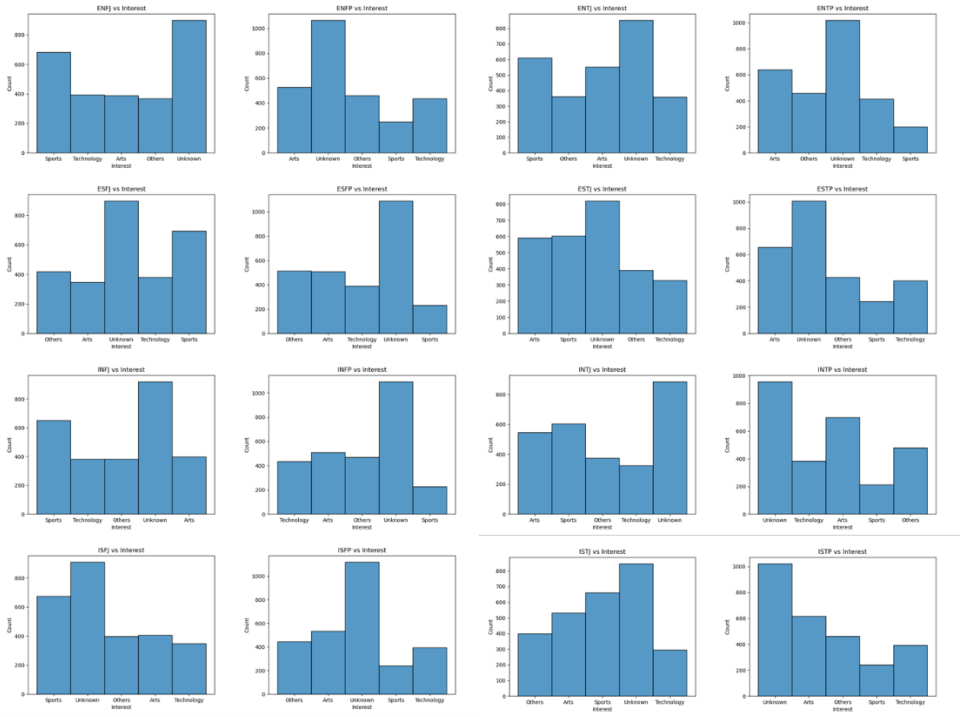

図 3-6　MBTI 別興味

図 3-6 より目立った法則性自体は確認できなかった。ただし、MBTI 別に何に興味があるかは分かれていることが分かった。

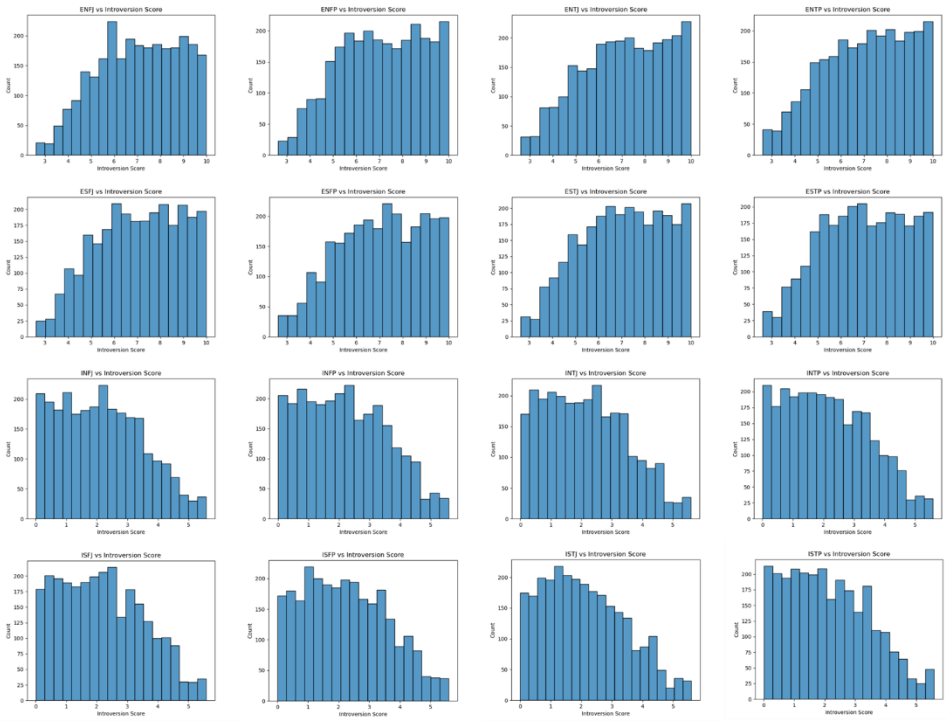

図 3-7　MBTI 別内向スコア

図 3-7 から基本的に E(外向的)は右肩上がりの結果だったのに対して、I(内向的)は右肩下がりの結果になった。

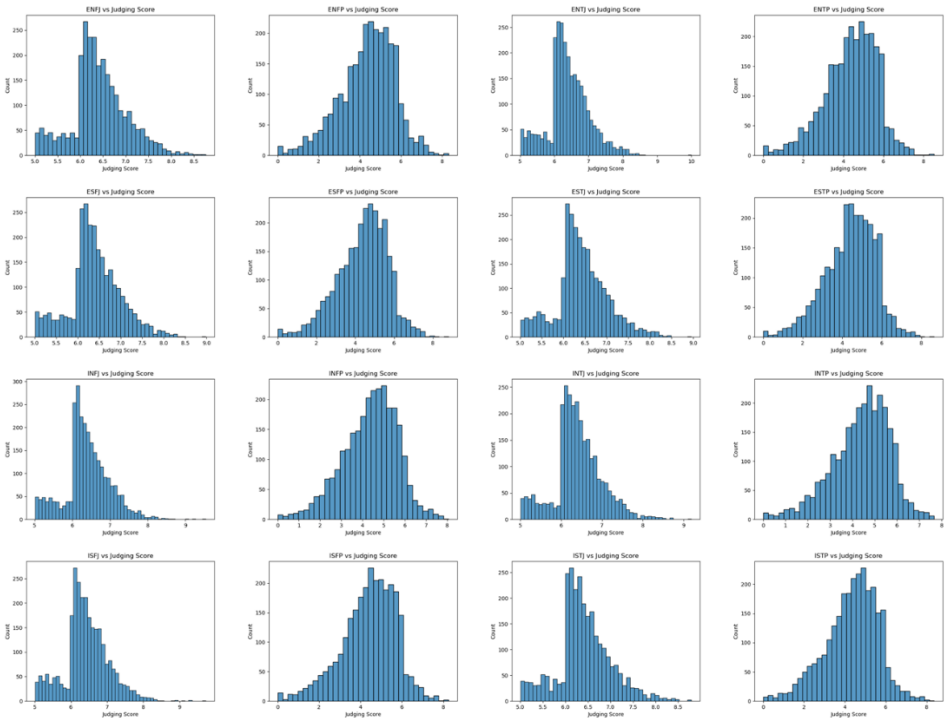

図 3-8　MBTI 別判断スコア

図 3-8 より J(判断型)はスコアが 6 辺りから急激に値が上昇しているおり、その後すぐ下降している。それに対して P(知覚型)はどれも近似した分布をしていることが分かる。

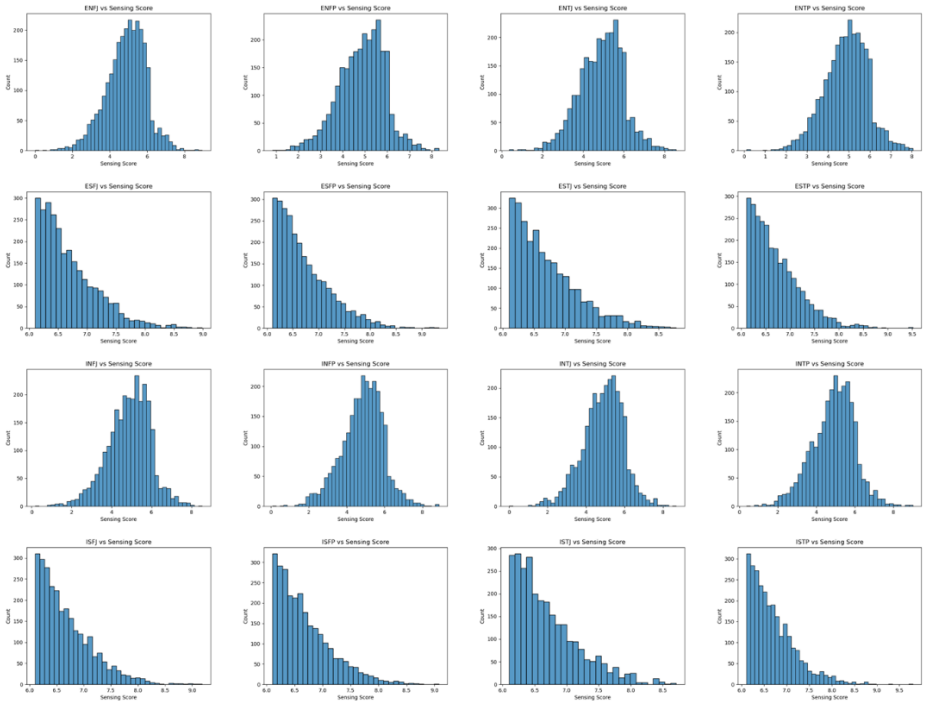

図 3-9　MBTI 別感覚スコア

図 3-9 から S(感覚型)はスコア 6 近くに最頻値がありそこから下降している。一方で N(直感型)は基本的に似た分布をしている。

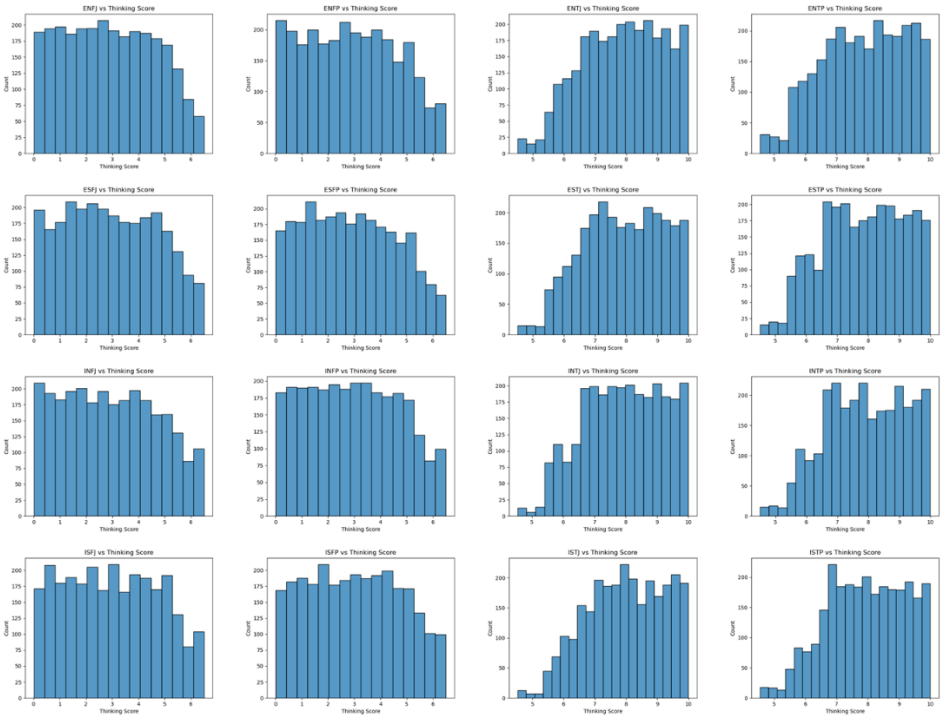

図 3-10　MBTI 別思考スコア

図 3-10 から T(思考型)はある地点からヒストグラムが右肩上がるがどこから上がるかが MBTI によって異なるのに対して F(感覚型)はある地点まで一様分布のような形をしてそこから右肩下がりのヒストグラムになっている。

4　クラスタについて検証

よりデータを深く分析するために主成分分析を用いて次元圧縮しデータを 2 次元に可視化した。この時に第一主成分と第二主成分の因子負荷量を分布とともにプロットする biplot の結果を図 4-1 に示す。

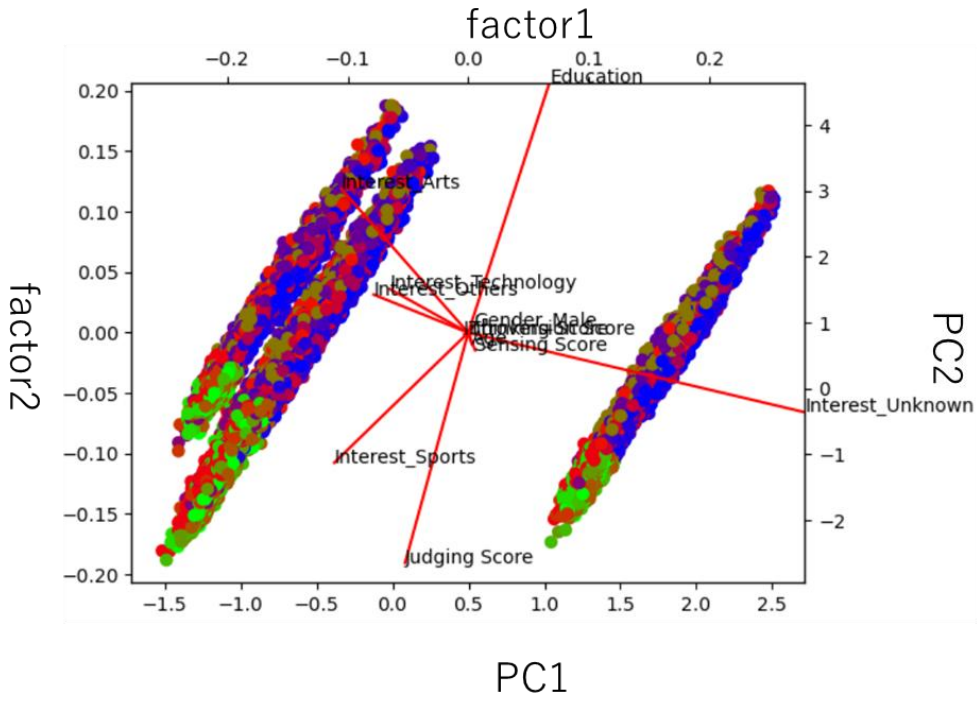

図 4-1　データセットの主成分分析と因子負荷量

なお主成分分析をするにあたり、尺度を統一するため標準化を行った。

また図 4-1 より 3 つのクラスタが確認できた。しかし MBTI についてはクラスタによって規則性(MBTI の値は色)が確認できないことが分かった。

biplot における因子負荷量の出力結果から横軸(第一主成分)は値が高いと Interest_Unknown が True(ダミー変数のため)になり Interest_Arts が False になる。また、因子負荷量から縦軸(第二主成分)は値が高いと教育の値が 1 になり、判断スコアが下がり Interest_Sports が False になる。

4.1　クラスタリング

この 3 つのクラスタについてクラスタリングを行いラベル付けを行う。

結果として行った主成分分析における第一主成分から第三主成分までを特徴量とし混合ガウスモデルを使用することで適切なクラスタリングができた(図 4-2)。

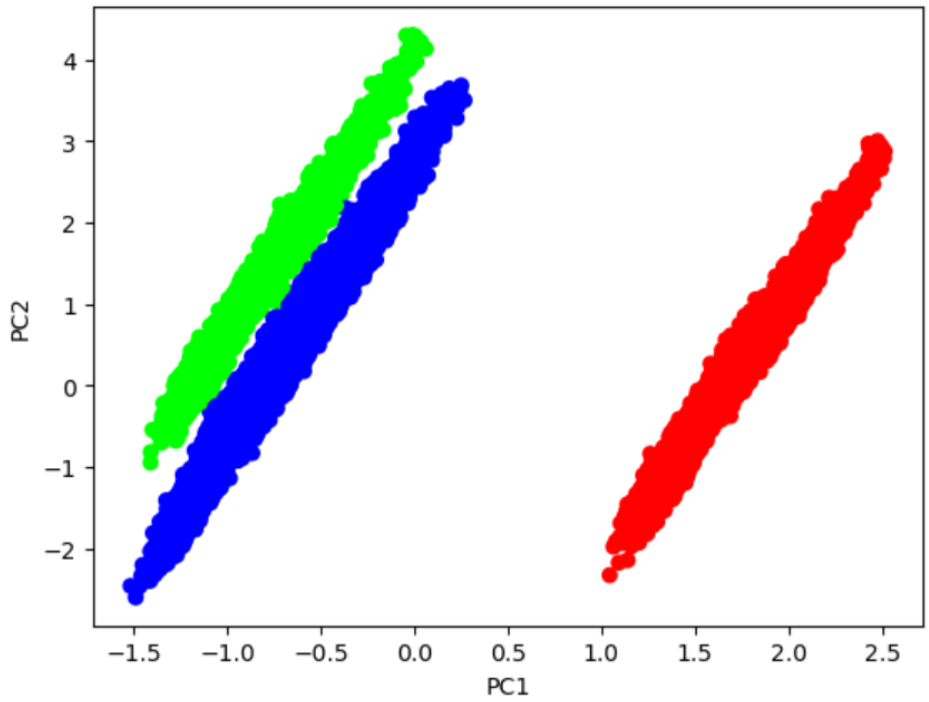

図 4-2　第一から第三主成分でクラスタリングをしてラベル付けを行った結果

4.2　第三主成分までの因子負荷量

図 4-2 のクラスタリングは第三主成分まで使用している。そこで第三主成分までを可視化したグラフを図 4-3 に示す。

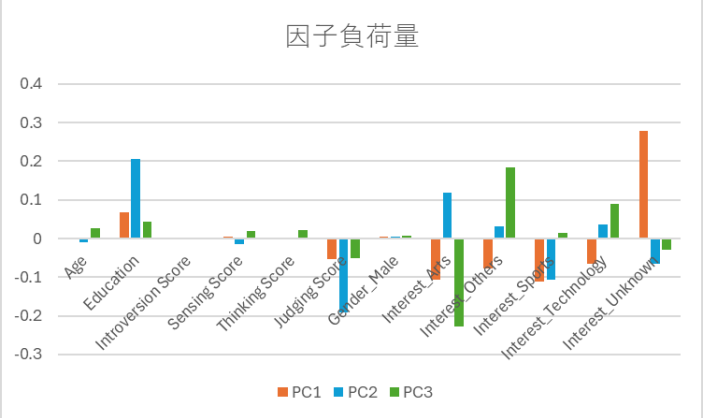

図 4-3　第三主成分までの因子負荷量

ここで、目立った第三主成分の因子負荷量については興味(芸術)が負に高く興味(他)と興味(科学技術)が正に高いことが分かる。

つまり第三主成分では値が正に大きいほど興味(他)・興味(科学技術)が True になり興味(芸術)が False になることが分かる。

4.3　クラスタリング結果を教師データにして分類

4.1 項で行ったクラスタリングによりラベル付けがされた。そこで元のデータセットを用いてクラスタの分類に寄与する特徴量を探す前段階としてクラスタリングされてできたラベルを教師データとして分類し精度を測定して分類寄与率の信ぴょう性を確かめる。

ここで訓練データ 50%でテストデータ 50%において得られた精度を図 4-4 に示す。

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10032
1	1.00	1.00	1.00	7629
2	1.00	1.00	1.00	4211
accuracy			1.00	21872
macro avg	1.00	1.00	1.00	21872
weighted avg	1.00	1.00	1.00	21872

図 4-4　クラスタリング結果を教師データとしたときの精度

図 4-4 より精度は 100%の結果になった。この結果としての信ぴょう性を元に計測した分類寄与率を表 4-1 に記す。

表 4-1　クラスタリング結果で測定した分類寄与率

	imp
Age	0.297986
Introversion Score	0.262228
Interest_Arts	0.122894
Interest_Unknown	0.122072
Sensing Score	0.104809
Thinking Score	0.045212
Judging Score	0.040691
Education	0.003288
Interest_Sports	0.000411
Interest_Others	0.000411
・	
・	
・	

測定した結果から上位から「年齢」「内向スコア」「興味(芸術)」「興味(不明)」「感覚スコア」「思考スコア」「判断スコア」「教育」という結果になった。

ここで上位になった特徴量のヒストグラムを図 4-5 から図 13 まで示す。

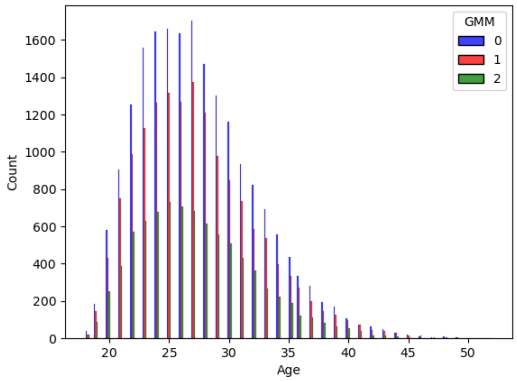

図 4-5　年齢のヒストグラム

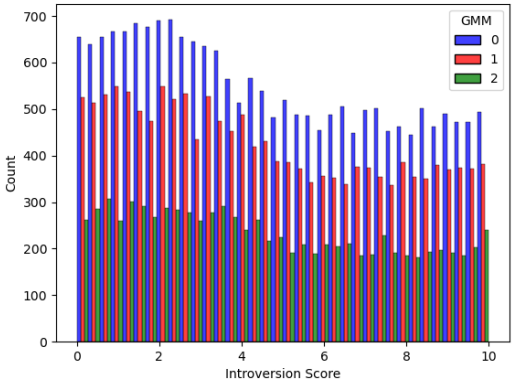

図 4-6　内向スコアのヒストグラム

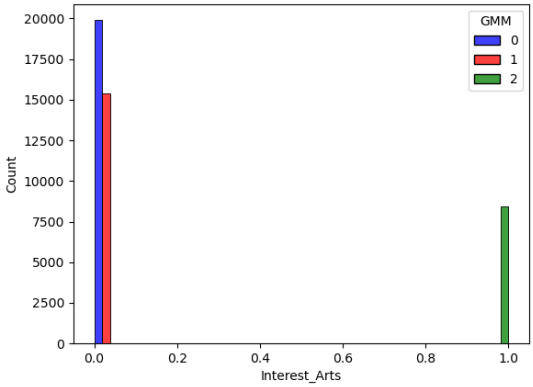

図 4-7　興味(芸術)のヒストグラム

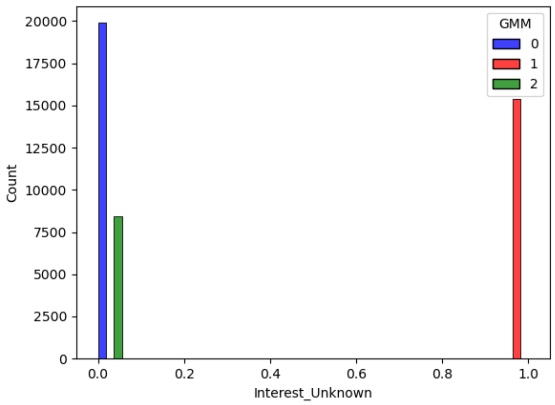

図 4-8　興味(不明)のヒストグラム

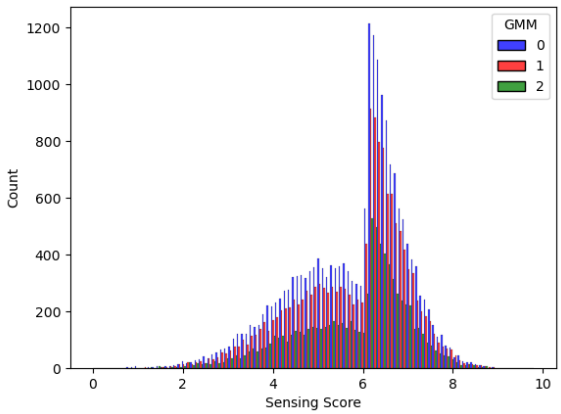

図 4-9　感覚スコアのヒストグラム

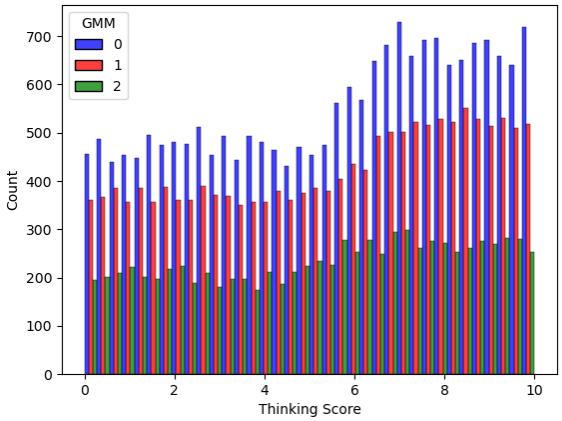

図 4-10　思考スコアのヒストグラム

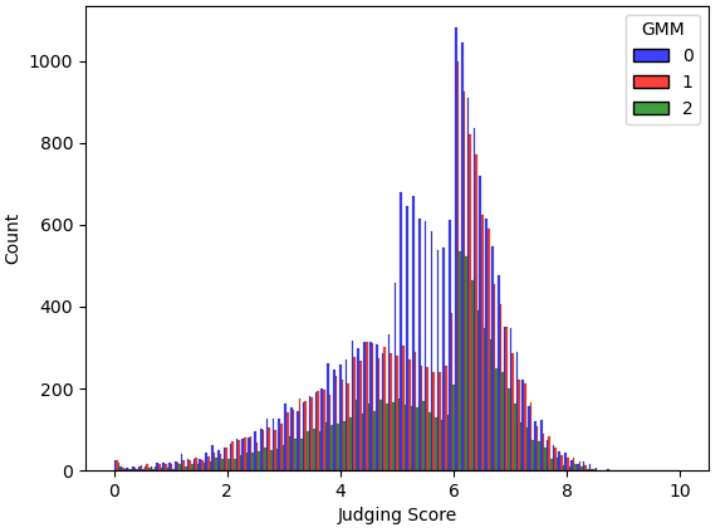

図 4-11　判断スコアのヒストグラム

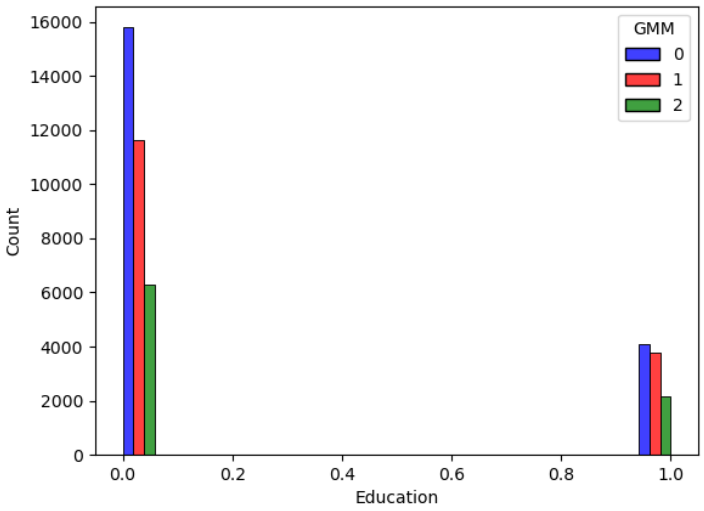

図 4-12　教育のヒストグラム

5　考察

5.1　MBTI データセットについて

図 3-3 から図 3-10 において、MBTI では陽性陰性の基準が明確にあり、それがデータセットに現れたことが表 3-2 における分類寄与率に現れたことが考えられる。

また、教育については P が大学院まで行き J が大学院まで行かない傾向があることがわかる。

E(外向的)の性別は S(感覚型)か N(直感的)かで男女比が逆転しており TJ(論理的思考)の場合、他の S か N と比率が逆転している。また、N の時は男性が多く、S の時は女性が多い。また、I(内向的)の性別は INFJ と ISFJ で同じ性別の多さは同じ程度の比率だが E と比べて男女比が大きく異なる。同様に、INFP と ISFP は性別が逆転しているが E と比べて男女比が大きく異なる。

5.2　クラスタリング

図 4-1 の biplot からクラスタ 2 つと 1 つを分ける特徴量は興味であることが分かる(特に興味(不明)、興味(芸術))。

図 4-2 からクラスタリングできた事の証左として図 5-1 にクラスタリングラベルごとの分布を示す。

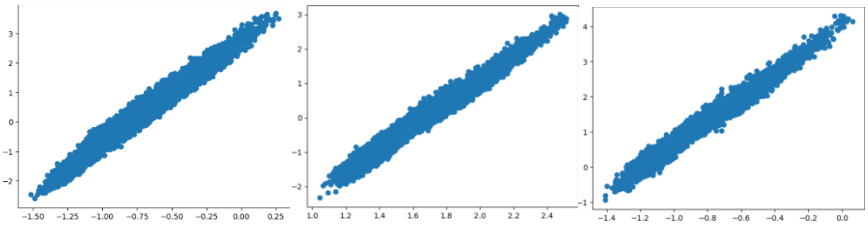

図 5-1　各ラベル毎のクラスタリング結果(左からラベル「0」「1」「2」)

この結果からクラスタリングは正常に行えた事が考えられる。

図 4-3 から各クラスタを特徴づけた因子としては「教育」「判断スコア」「興味(芸術)」「趣味(他)」「興味(スポーツ)」「興味(科学技術)」「興味(不明)」等が挙がっている。また、分類寄与率とヒストグラムから興味軸は特に「興味(芸術)」「興味(不明)」がクラスタを分ける要因になっていることが分かる。その他では判断スコアと教育が若干分布が異なっており、明確に分かれるのは「興味(芸術)」「興味(不明)」の 2 つであることからこのクラスタは「興味(芸術)」「興味(不明)」を分けるものとなっており判断スコアと教育は特徴量を補完していると考えられる。

6　まとめ

MBTI の 93 個の質問が完璧だと仮定したとき、今回の 8 個の特徴量では代表値の中で最頻値を採用するとしたときに 10%程度精度が落ちてしまうことが分かった。

また、MBTI とは関係ない事項(年齢・性別)においては性別は外向的(E)の時に関係しやすいことが分かったが内向的(I)の時は性別において規則性は確認できず、年齢については細かい分布の違いはあれど目視では分類に寄与するほどの大きな違いが確認できなかった。

また、趣味趣向については分類においては寄与が確認できなかったが MBTI が毎に異なっており、これについて法則性は確認できなかった。

次に主成分分析をした結果 3 つのクラスタが現れ、第三主成分まで特徴量としてクラスタリングに有効であることが分かった。

これについて、因子負荷量と GMM で作られたクラス別ヒストグラムからは 3 つのクラスタは興味についてがメインで判断スコアと教育がデータを補完していることが分かった。