7G-04

統計的解析を用いたアンケートデータからの特徴抽出手法の検討

岡本　大輝†　　後藤　淳†

NHK 放送技術研究所†

1. はじめに

近年のコンピュータ性能の向上に伴い，多量の計算を要する高度な分析・予測手法が多数提案されている．特に，ディープニューラルネットワークに代表される機械学習等，多量の並列計算を行う分析・予測技術の発展は目覚ましく，活用事例は近年急速に増えている[1]．こうした技術は一般的に，学習データの充実した画像や音声など大規模なデータに適している．

一方，取材やマーケティングの領域ではインターネットを使った低コスト調査が普及してきており，世論調査においてもインターネットを活用した調査に関心が注がれている[2]．これらの調査・分析では一般的に，アンケート等によって得られた数千サンプル程度の小規模データを扱うが，機械学習の精度を高めるには不十分な規模になり易い．また，予測の精度だけではなく理由を付した分析も求められるが，予測精度の向上を目的とした機械学習によって導き出される条件やルールが複雑になりやすく，一般的には理解できないことが多い．

結果として，アンケートデータを対象とした分析において，機械学習が用いられる事例は限られる．ＮＨＫにおいても，番組制作の過程でアンケート調査を行うことがあるが，十分な経験と知識を持つ番組制作者の地道な分析に頼る必要があった．

今回，インターネットを用いたアンケート調査のデータを対象として，特定の回答をした回答者を”目的群”として，その特徴を抽出する実験を行う．具体的には，アンケートに含まれやすく分析も容易である選択式の設問に着目し，回答の分布をもとに回答者を多次元空間にマッピングする．多次元空間の中で目的群の特徴を示す特徴ベクトルを算出し，ベクトル成分が示す目的群の特徴について考察する．特徴ベクトルの算出手法として，算出結果が人に理解されやすいとされる主成分分析に着想を得た手法を提案する．また，同様の条件で機械学習によって特徴ベクトルを算出し，提案手法で算出されたベクトルと比較する．さらに，特徴ベクトルと回答者の座標ベクトルとの内積をスコアとして AUC を算出し，各特徴ベクトルの分類性能についても考察する．

2. 多次元空間へのデータマッピング

最初に，分析の対象とする目的群の条件を決定する．実験では，あらかじめアンケートデータの回答者を開発セットとテストセットに分けた上で，開発セットに含まれる目的群以外の回答者群である”非目的群”を抽出し，非目的群の設問の内容と回答を参照し，回答者を多次元空間にマッピングするルールを作成する．以下，選択式設問の設問内容によって異なる 2 種類のルールについて説明する．

Examination of Statistical Feature Extraction Methods for Questionnaire Data
†NHK Science and Research Laboratories

・程度を問う選択肢を有する設問 q の場合

『強くそう思う』『そう思う』など，設問に対して回答者の意見や行動などの程度を問う選択肢から，１つを選択して回答するタイプの設問である．

まず，非目的群回答者の全回答からヒストグラム $h_{q,c}$ を作る．c は選択肢ごとに対応する非負整数で，選択肢に示される程度の順に対応する．回答者の分布が正規分布に従うと仮定して，正規分布の累積分布関数とヒストグラムを対応させ，回答者の座標が確率変数の期待値 $x_{q,c}$ と等しくなる一つの次元を定義する．c より低い程度の回答をした回答者の総数を $h_{q,c}^{(-)}$，高い程度の回答をした回答者の総数を $h_{q,c}^{(+)}$，回答者の総数を N_q として，

$$a_{q,c} = erf^{-1}\left\{\left(h_{q,c}^{(-)} - h_{q,c}^{(+)} - h_{q,c}\right)/N_q\right\}$$
$$b_{q,c} = erf^{-1}\left\{\left(h_{q,c}^{(-)} - h_{q,c}^{(+)} + h_{q,c}\right)/N_q\right\}$$
$$x_{q,c} = \frac{N_q}{h_{q,c}}\left\{exp\left(-{a_{q,c}}^2\right) - exp\left(-{b_{q,c}}^2\right)\right\}$$

となる．ここで，erf^{-1} は誤差逆関数である．

・複数回答問題および程度を問わない択一問題の場合

『この中から当てはまるものを[一つ/全て]選んで下さい』と問うタイプの設問である．

選択肢のそれぞれを一つの次元として，選択されたか否かで選択肢ごとにヒストグラムを作り，程度を問う択一問題と同様の方法でマッピングする．すなわち，ヒストグラム $h_{q,c}$ において，$c \in \{0,1\}$ となる．

上記 2 種類のルールから，回答者を r 次元の空間にマッピングするルールを作成する．空間の各次元はそれぞれ，もとのアンケートの設問や選択肢に対応する．なお，マッピングルールは開発セットの非目的群のみを用いて作成するが，開発セットの目的群やテストセットのデータに対しても，同じマッピングルールを適用する．

3. 特徴ベクトルの算出

マッピングルールを開発セットのデータに適用し，目的群の回答者の座標ベクトルを各行に有する行列 X_p と，同じく非目的群の行列 X_n を得る．これらの行列に対し，以下に述べる 3 つの手法を適用し，目的群の特徴ベクトルをそれぞれ算出する．特徴ベクトルはいずれも 2 節で定義した r 次元の成分を有する．特徴ベクトルの成分のうち，絶対値の大きな成分に対応する設問や選択肢が目的群に特徴的であり，成分の正負が回答傾向に対応する．なお，マッピングルールは非目的群の回答分布を正規分布と仮定して作成しているいるため，X_n の列方向に平均を取った行ベクトル $E(X)$ は，全ての成分が 0 になる．

3.1. 提案手法(Prop)

対称行列 $X_n^T X_n$ の固有値と固有ベクトルをそれぞれ

λ , v とする．$X_n^T X_n$ は半正定値行列なので，固有値 λ は全て正の値となり，その種類数，すなわち $X_n^T X_n$ のランクを s とする．固有値を対角成分にとる対角行列を $\Lambda(\in R^{s\times s})$ ，Λ の各固有値に対応した固有ベクトルを行の成分とする行列を $V(\in R^{s\times r})$ として，行列 $F = \Lambda^{-1/2}V$ を計算する．ここで，X_n を線形変換したデータ $X_n F^T = Y_n\,(\in R^{N\times s})$ は，s 次元空間のいかなる方向に対しても分散が 1 となる分布をとる．すなわち，X_n の多次元正規分布を標準化する作用を有する．

X_p を線形変換したデータ $Y_p = X_p F^T$ を計算し，対称行列 $Y_p^T Y_p$ の固有値と固有ベクトルを求める．得られた固有ベクトルを $u(\in R^s)$ として，要素数 r のベクトル uF を複数の固有ベクトルごとに求める．

3.2. 提案手法を簡略化した手法(UNmlz)

提案手法において線形変換行列 F を単位行列に置き換え，もとの目的群の分布 X_p に対して直接特徴ベクトル抽出を行う．線形変換行列 F によって非目的群の分布が標準化されるが，それによって期待される目的群に固有の特徴の強調効果を検証する．

3.3. 線形 SVM のサポートベクトル(SVM)

X_p を正例，X_n を負例として線形 SVM を適用し，得られたサポートベクトル(以下，SV)を特徴ベクトルと同様に検証する．提案手法とその簡易版においては，互いに直交した特徴ベクトルが複数求められるため，それを模して以下のように SV を複数求める．

- X_p と X_n を線形 SVM に適用して SV を得る
- SV 方向の分布成分をそれぞれ縮退させる
- 縮退させたデータを再び線形 SVM に適用する

上記の手順を繰り返し，互いに直交した SV を複数求める．

4. 評価実験

4.1. 特徴抽出を行うデータセット

2020 年 12 月 10 日から 2 日間，web アンケートに回答した 1,330 名分のデータを対象とした．設問の内容は多岐に渡り，接触メディア，食生活，コロナウイルス，東京オリンピック，環境問題などとなる．実験では『東京オリンピックは楽しみですか』という設問に対して『とても楽しみ』『どちらかというと楽しみ』と答えた 534 名の回答者を目的群として，目的群の回答傾向を抽出する実験を行った．設問数は 24，基本プロファイルも含めた合計特徴量数 r は 88 であった．

4.2. 実験設定

データを 5 分割し，1 セットずつ順にテストセットとする 5 回の試行を，分割方法を変えながら 10 ループ，計 50 回の試行を行った．各試行で固有値の大きな(SV の場合は出現順の早い)上位 3 つのベクトルを集め，手法ごとに 150 のベクトルを作成した．150 のベクトル同士のコサイン類似度を計算し，他のベクトルとの類似度の総和が大きなベクトルを基点に，コサイン類似度 0.2 以上を閾値としてクラスを生成した．40 以上のベクトルが属し，属するベクトルの平均固有値の大きな(平均出現順の早い)2 クラスに属するベクトルの平均を Prop_1，Prop_2，UNmlz_1，UNmlz_2，SVM_1，SVM_2 として検証した．

手法 (AUC)	絶対値の大きな成分に対応する傾向
Prop_1 (0.6826)	五輪競技を PV 観戦したい
	普段からスポーツを PV 観戦する
	五輪競技をリアルタイム視聴したい
Prop_2 (0.7106)	五輪競技を生で観戦したい
	五輪競技を PV 観戦したいとは思わない
	普段からスポーツ観戦に出かける
UNmlz_1 (0.8041)	五輪は暮らしに良い影響があると思う
	五輪競技を PV 観戦したい
	普段からスポーツを PV 観戦する
Unmlz_2 (0.6943)	五輪競技を PV 観戦したいとは思わない
	普段からスポーツを PV 観戦しない
	五輪競技をリアルタイム視聴したい
SVM_1 (0.8316)	五輪競技をリアルタイム視聴したい
	五輪は経済に良い影響があると思う
	リアルタイムに拘らず五輪競技を見たい
SVM_2 (0.8150)	五輪競技をリアルタイム視聴したい
	五輪は暮らしに良い影響があると思う
	五輪は経済に良い影響があると思う

表：『オリンピックを楽しみにする人』の回答傾向と特徴ベクトルの AUC

※ “PV”はパブリックビューイング

4.3. 実験結果と考察

各平均ベクトルについて，多次元空間にマッピングしたテストデータの座標との内積をスコアとして作成したランキングの AUC と，絶対値の大きなベクトル成分と対応する回答傾向をまとめたものを表に示す．

提案手法(Prop)から得られたベクトルは，いずれもスポーツ観戦行動についての設問に対応する成分が大きな値を取り，「パブリックビューイングで観戦する人」と「実際に会場で観戦する人」を示していると考えられる．簡略化した提案手法(UNmlz)ではその特徴が薄れていた．一方，SVM から得られる SV は最も高い AUC を得られたが，ベクトル成分と対応する設問は行動と意識を問うものが混在していた．また，次元縮退を経て得られた 2 つのベクトル同士の中身が似ていた．目的群に内在する複数の異なる回答傾向が混合されることで，分類性能が向上する代わりに傾向抽出が困難になったと考えられる．

5. まとめ

本稿では，アンケート調査で特定の回答をした回答者を目的群として，目的群以外の回答者群をもとに回答者を多次元空間へマッピングするルールを作り，目的群を多次元空間中の固有ベクトルで特徴づける手法を提案した．実験では，目的群の回答傾向を分析して互いに異なる傾向の特徴を抽出できることを示した．今後，主観評価も実施し，アンケートデータを分析する番組制作者を補助するツールとしての機能を高めていく．

参考文献

[1]. 根岸 他，“ディープラーニング活用事例と使いこなしの勘所”，情報処理 vol.59 No11，2018，

[2]. 放送文化研究所，“住民基本台帳からの無作為抽出による WEB 世論調査の検証②”，放送研究と調査 2018 年 9 月号，2018，