7U-04

マルチモーダル対話におけるクロスコーパスでの心象推定のための特徴量選択

久保 祐喜 * 羅 兆傑 * 武田 龍 * 駒谷 和範 *
* 大阪大学 産業科学研究所

1. はじめに

対話環境は収録に用いるデバイスや周囲の環境によって多様に変化する．ユーザの発話からユーザの心象を推定する学習モデルは多様な場面で動作することが望ましい．マルチモーダル対話が収録されている Hazumi[1] では，対面やリモートのような複数の環境での収録が存在する．図 1 のように，訓練データとテストデータで収録環境が異なることで精度が低下する場合がある．この精度の低下には，クロスコーパスでの心象推定において，悪影響を及ぼす特徴量が存在することによると考えられる．

ユーザの発話から興味度や心象を推定する手法は複数存在する．特定の収録環境の下で，ユーザの発話を基にユーザの興味度や心象等を推定するものとしては，マルチモーダル情報を用いた手法 [2, 3, 4] が提案された．本研究では，複数の収録環境での心象推定を行う．また，クロスコーパスでの感情推定について，Gao ら [5] はコーパスの情報による特徴量の表現の正規化を伴う学習器を提案した．本研究では，特徴量の正規化以外の着目点として，心象推定に悪影響を及ぼす特徴量が存在しているかを確認することを注目し，特徴量選択を用いる．

2. 特徴量選択の適用

図 2 のように，音声や顔画像について，特徴量の一部を選択し，心象推定を行う．特徴量選択は，Wrapper 法および Filter 法を用いた [6]．Wrapper 法を用いる目的は，対話環境が変化していない場合と変化した場合の精度の差が小さくなるかを確認することである．Filter 法を用いる目的は，心象と相関の高い特徴量を選択した場合，Wrapper 法と近い結果が得られているかを確認することである．

2.1 Wrapper 法に基づく特徴量選択

Wrapper 法の前向き探索を行うことで，クロスコーパスでの心象推定で有効な特徴量の組み合わせを探索する．訓練データとテストデータを設定し，特徴量を選択しながら検証を行うことで，最も心象推定の精度が高い特徴量の組み合わせを探索する．訓練データとテストデータの両方を用いて特徴量の組み合わせを探索しているため，この結果をオラクルとする．選択する特徴量の集合 S（初期では空集合），元の特徴量の集合 F を準備する．以下の 1 と 2 の処理を全ての特徴量が選択されるまで繰り返し，精度が最大のときの特徴量の組み合わせを Wrapper 法の結果（S）として使用する．

1. F から特徴量を 1 つ候補として取り出し，S に加えた場合の精度を求める．
2. S に加えたときに最も良い結果となる特徴量 1 つを F から S へ移動する．

Feature Selection for Cross-Corpus Sentiment Analysis in Multimodal Dialogues: *Yuki Kubo, Zhaojie Luo, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

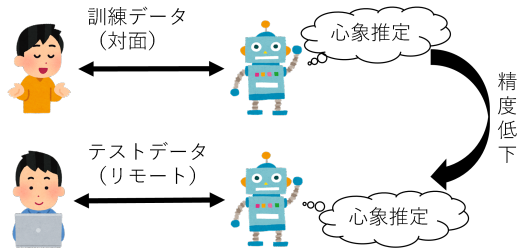

図 1: 対話環境の変化による心象推定の精度低下の例．

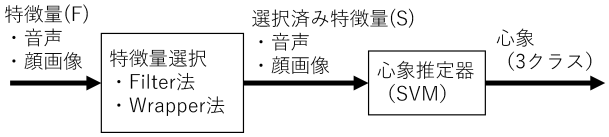

図 2: 特徴量選択を伴う心象推定のデータの流れ．

2.2 Filter 法に基づく特徴量選択

訓練データ内での特徴量と心象の間の相関と交差検証に基づき，特徴量を選択する．以下の処理を $K = 1, \ldots$ で繰り返し，交差検証の精度が最大のときの特徴量の組み合わせを Filter 法の結果（S）として使用する．

1. 訓練データに対して SelectKbest[2]を適用し，F から心象との相関の高い特徴量を K 個取得する．
2. 訓練データ内で一人抜き交差検証を行い，選択された K 個の特徴量の組み合わせでの精度を求める．

3. 実験

3.1 実験データ

本研究で使用したデータセットは Hazumi である[3]．Hazumi では Wizard of Oz 方式の対話システムとユーザによるマルチモーダル対話が収録されている．第三者が付与したユーザの心象を 3 クラスにしたラベルを使用する．本研究で使用したラベルの分布は表 1 に示す通りである．Hazumi には，対面で収録したもの（1712, 1902,

表 1: 各収録版での正解ラベルの分布

収録環境	バージョン	低群	中群	高群
対面	1712	419	855	987
	1902	175	999	1163
	1911	178	881	1380
リモート	2010	91	1049	1658
	2012	245	1511	3578
	2105	49	499	1682

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html

[3] 本研究では，https://github.com/ouktlab にて公開されているダンプファイルを使用した．

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.

表 2: 訓練データとテストデータに用いた収録環境単位での正解率の平均．バージョン単位で訓練データとテストデータを割り当てて実験し，訓練データとテストデータでの収録環境ごとにグループ分けをして正解率の平均をとった．

モダリティ	訓練	テスト	majority	ベースライン	Wrapper	Filter
音声	リモート	リモート	0.67	0.81	0.85	0.81
	対面	リモート	0.67	0.43	0.82	0.69
	対面	対面	0.50	0.63	0.71	0.63
	リモート	対面	0.50	0.51	0.70	0.56
顔画像	リモート	リモート	0.67	0.73	0.76	0.72
	対面	リモート	0.67	0.29	0.75	0.44
	対面	対面	0.50	0.61	0.66	0.63
	リモート	対面	0.50	0.55	0.64	0.54

1911：以降まとめて対面版）と Zoom を使用して収録したもの（2010，2012，2105：以降まとめてリモート版）が存在する．

音声特徴量はユーザ発話から openSMILE を用いて抽出された 384 次元の特徴量である．顔画像特徴量はユーザの顔画像から OpenFace を用いて出力した数値について，ランドマークの速度に関する演算や Action Unit の平均によって構成される 66 次元の特徴量である．

3.2 実験設定

音声特徴量と顔画像特徴量について実験を行った．訓練データとテストデータとして対面版とリモート版のバージョン群から 1 つずつデータセットを使用し，心象推定の正解率を求めた．訓練データとテストデータで同じバージョンを使用しないようにした．正解率について，訓練データとテストデータでの収録環境ごとにグループ分けをした．このグループ分けした結果群ごとに，平均を取ったものを表 2 に示した．

表 1 での最も個数の多いラベル（高群）の割合をバージョン単位で算出して収録環境単位で平均したものをマジョリティーベースライン（majority）とした．心象推定器の学習結果について，特徴量選択を行わずに全特徴量を入力した際の結果（ベースライン），及び Wrapper 法を適用した際の結果，Filter 法を適用した際の結果を示した．

心象推定器として，SVM を使用した．カーネルは RBF カーネルを使用し，$C=1, \gamma=\frac{1}{feature_length}$（$feature_length$ は SVM へ入力する特徴量の次元数を表す）とした．訓練と検証で用いる特徴量は訓練データの平均と分散によって標準化を行った．

3.3 実験結果

実験結果を表 2 に示す．Wrapper 法と Filter 法での結果について，それぞれ考察を行う．

先ず，Wrapper 法について述べる．あるテストデータの収録環境に対して，訓練データの収録環境が同じ場合と異なる場合の正解率の差を比較した．Wrapper 法での結果について，収録環境が訓練データとテストデータで異なっている場合の心象推定の正解率の低下が，音声と顔画像の両方で 0.03 以下になった．これは，ベースラインでの収録環境が異なっている場合での正解率の低下よりも小さい結果となった．特徴量選択によって差が小さくなることにより，異なる対話環境で運用する場合でも劣化が抑えられていることが確認された．これにより，特徴量の中にクロスコーパスでの心象推定に悪影響を及ぼすものが含まれていることで精度低下が発生していることが確認された．

次に，Filter 法について述べる．訓練データとテストデータの収録環境が異なる場合での正解率について，majority やベースラインと比較する．音声ではベースラインと比べて正解率が 0.05 以上向上した．顔画像では，対面版で訓練しリモート版でテストした場合の正解率はベースラインより高いが，majority より低い値となった．これにより，音声では，Filter 法による特徴量選択が有効であると考えられる．

4. おわりに

今回，収録環境の変化による心象推定の精度の低下を課題とした．クロスコーパスでの心象推定において有効な特徴量の存在を確認するために Wrapper 法と Filter 法を適用した．Wrapper 法では，訓練データとテストデータの収録環境が異なる場合と同じ場合の正解率の差が 0.03 以下でとなることを確認した．これにより，クロスコーパスでの心象推定に悪影響を及ぼす特徴量が存在することが確認された．Filter 法では，訓練データ内の心象と相関に基づいて特徴量を選択することで精度の低下を一定量抑えたことを音声特徴量について確認した．

今回は，音声と顔画像それぞれ別個に検証を行ったが，音声と顔画像を組み合わせることが有用であると考えられる．また，今回の問題は標準化の方法や学習モデルの設定にも依存するため，複数の実験条件での検証が必要であると考えられる．

参考文献

[1] 駒谷和範. マルチモーダル対話コーパスの設計と公開. 日本音響学会誌, Vol. 78, No. 5, pp. 265–270, 2022.

[2] 松本紗規子ほか. 雑談対話におけるマルチモーダル情報を統合した興味判定手法. 人工知能学会研究会資料 言語・音声理解と対話処理研究会, Vol. 84, pp. 86–87, 2018.

[3] 西本遥人ほか. 対話におけるマルチモーダル情報を用いたユーザの興味の有無の推定. 人工知能学会第 32 回全国大会, 2018.

[4] Yuki Hirano, et al. Multitask prediction of exchange-level annotations for multimodal dialogue systems. In *2019 International Conference on Multimodal Interaction*, pp. 85–94, 2019.

[5] Yuan Gao, et al. Domain-invariant feature learning for cross corpus speech emotion recognition. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6427–6431, 2022.

[6] 元田浩ほか. データマイニングの基礎. 株式会社オーム社, 2006.

Copyright ©2023 Information Processing Society of Japan.
All Rights Reserved.