

プログラミング時の自己問題解決を目的とした困惑検出を介入契機とする アヒル型ロボット

倉田 寛大[†] 白神健瑠^{††} 井上希凛^{††} 森岡怜司^{†††} 大澤隼也^{††}

辻 愛里^{††††} 藤波 香織^{†,††††}

[†] 東京農工大学 大学院 生物システム応用科学府 生物機能システム科学専攻

^{††} 三菱電機株式会社 情報技術総合研究所

^{†††} 三菱電機株式会社 静岡製作所

^{††††} 東京農工大学 大学院 工学研究院 先端情報科学部門

1 はじめに

近年、IT 人材の不足が深刻化する中で、人手をかけずにシステムを活用して、さまざまなレベルのプログラマを幅広く支援する取り組みには重要な意義がある。我々はこれまで、ユーザの発話に応じて相槌を打つアヒル人形型システム（以下、uh-huh duck）を実装し、情報整理の支援における相槌機能の有効性を明らかにした。その一方で、作業者が自発的に困惑状態を報告し、それを契機として支援するシステムであったために、支援すべきタイミングを逃す事例が見られた [1]。そこで、困惑検出時にシステム契機で支援を実行することを目指して、既報 [2] にて困惑状態推定モデルの作成を試みた。本稿では、リアルタイムで動作する推定モデルを構築し、uh-huh duck に組み込んで評価した結果を報告する。

2 uh-huh duck システム

2.1 概要

提案システムの概要を図 1 に示す。ユーザからデバイスを介して、打鍵情報や目の動き、感情を取得し、それぞれ訓練済みの分類器にて、困惑度合を 5 段階で推定する。これらの推定結果を一つのデータベースに記録し、定期的に加重平均にて 1 つの結果に統合し、uh-huh duck の相槌制御部に送信する。推定結果に応じて uh-huh duck は起動する。本システムはマイクを介してユーザの発話の切れ目を検出すると、モータとスピーカによって相槌の動作と音声を実現する。以下、推定モデルの作成と組み込みの過程について説明する。

2.2 困惑状態推定手法

複数のデータ源から個別に分類器を作成し、それぞれの推定結果を統合するアンサンブル分類の形を採用する。以下、個別の分類器作成から評価データの収集、および分類器の性能向上を図った内容を記す。

2.2.1 個別の分類器作成

性能向上が最大限見込まれる個人内での訓練を実施し、個人ごとに 3 つの分類器を作成した。既報 [2] で収集したデータのうち、図 1 にある 3 種類の生データをを用いて特徴量を計算し、訓練データとした。なお、検証の結果から、打鍵情報および目の動作に関しては、それぞれスライディングウィンドウを用いた特徴量算出を行い、そのウィンドウサイズは 25 秒および 6 秒とし

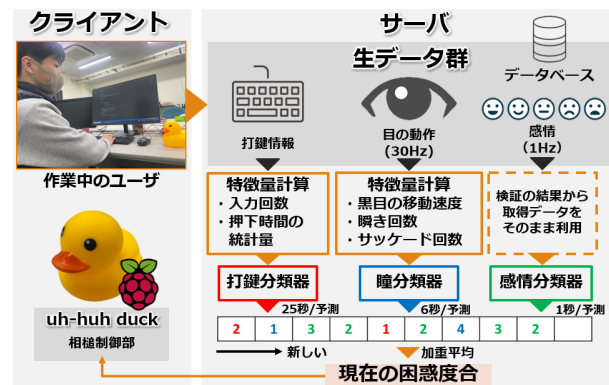


図 1: システム概要図

た。一方、感情データ [3] については、サンプリングされた値を直接利用した。

2.2.2 評価データの収集

個人内で訓練した分類器の実使用時の性能を評価するため、既報 [2] と同様の環境で新たに評価データを収集した。被験者は、情報系の学部及び大学院に所属する平均 22.8 歳 ± 1.8 の 12 名（男性：5 名、女性：7 名）である。デバイスを卓上に配置した状態でプログラミングタスクを与え、1 分間隔でポップアップ表示されるアンケートから、現在の困惑度合と、誰かに話すことで思考を整理したいか否かを回答させた。これにより、3.1 節における評価データを収集するとともに、3.2 節の実験で使用する uh-huh duck 起動の閾値を調査した。

2.2.3 各分類器のチューニング

分類器はマクロ平均 F 値が最大となるようハイパーパラメータの調整を行った。分類モデルには Random Forest を採用し、弱分類器の数、木の最大の深さ、分割基準の 3 つのハイパーパラメータを最適化した。なお、2.2.2 項で収集したデータを評価に用いた。

2.3 uh-huh duck への組み込み

サーバから推定困惑度合を受け取り、その度合に応じてシステムを起動させるか否かを判定する。判定には、2.2.2 項で収集した、“誰かに話して整理したい”項目の、困惑度合に対する分布を用いることとした。具体的に、困惑度合が 5 段階中の 5 であった際に“話して整理したい”と回答した割合が 80% であった場合、推定の困惑度合が 5 と出力された際には、80% の確率でシステムを起動させる仕様とした。

また、既報 [1] と同様に、スイッチの押下で起動する仕様は残しつつ、音声が入力検出されない場合は、プログラムを停止するよう設計した。

3 評価

3.1 個別の分類器の出力統合方法

各分類器が推定した困惑度合は、1 つのデータベースに集約される。しかし、各分類器で最適なウィンド

A duck robot for triggering an intervention for self-problem solving via confusion detection during programming

[†] Kandai KURATA ^{††} Takeru SHIRAGA ^{††} Kirin INOUE ^{†††} Reiji MORIOKA ^{††} Shunya OSAWA ^{††††} Airi TSUJI ^{†,††††} Kaori FUJINAMI

[†] Graduate School of Bio-Applications and Systems Engineering, Tokyo University of Agriculture and Technology, ^{††} Mitsubishi Electric Corporation, Information Technology R&D Center

^{†††} Mitsubishi Electric Corporation, Shizuoka Works

^{††††} Division of Advanced Information Technology and Computer Science, Tokyo University of Agriculture and Technology

ウサイズは異なるため、一定区間内の推定結果の数は分類器によって大きく異なる。加えて、分類器に性能差があったり、結果の統合方法も複数考えられるため、以下の項目に関して検証を行った。

- 直近 30 秒の推定結果全てを使用する場合と各分類器の最新の推定結果のみを使用する場合
- 分類器の性能差を考慮して重みを付ける場合と付けない場合
- 多数決による場合と平均による場合

これらを個人ごとに検証し、マクロ平均をとった結果を表 1 に示す。なお、性能差を考慮する条件では、各分類器のマクロ平均 F 値を正規化した値を重みとし、多数決の場合は、推定されたクラスに分類器の重みを加算、平均の場合は、各クラスの事後確率に重みを乗算した。なお、多数決で推定結果が複数のクラスに分かれた場合は、平均をとった。

表 1: 統合方法の検証結果 (マクロ平均 F 値)

直近 30 秒分全て使用				最新の結果のみ使用			
性能差考慮		性能差無視		性能差考慮		性能差無視	
多数決	平均	多数決	平均	多数決	平均	多数決	平均
0.236	0.239	0.237	0.232	0.229	0.218	0.237	0.224

次節の実験では、最も高い F 値を示した、直近 30 秒分の推定結果を全て使用し、分類器の性能差に応じて加重平均をとる手法を採用する。なお、統合前の各分類器のマクロ平均 F 値は、打鍵分類器が 0.234、瞳分類器が 0.238、感情分類器が 0.224 であり、いずれも統合後のスコアに劣るため、統合の有効性が確認できる。

さらに、提案手法の有効性を検証するために、scikit-learn (バージョン 1.2.2) の DummyClassifier を用いてクラスごとのデータ数の分布に従うランダムな分類器を作成し、同じ評価データに対して推定を行った。その結果、マクロ平均 F 値が 0.163 となった。この結果と提案手法との間で Wilcoxon の符号順位検定を行ったところ、有意水準 5% で有意差が認められた。このことから、提案手法は少なくともランダムな分類器より有意に高い F 値で困惑度合を推定できることがわかる。

なお、本稿では 3 種類の分類器の結果を統合しているが、統合する分類器の数に制限はない。そのため、将来的には異なるデータ源から構築した分類器を追加して活用することも可能である。

3.2 システム利用による思考整理効果

3.2.1 実験設定

システム起動契機と思考整理効果の関係を調査するため、以下の 3 条件で実験を行った。

- ユーザ契機：スイッチ押下時のみ起動
- システム契機：困惑検出時およびスイッチ押下時に起動
- ランダム：ランダムなタイミングおよびスイッチ押下時に起動

全システムにおける返答内容は、「何か悩んでいるの？」等、実験者が Wizard of Oz 法にてその都度適切に決定した。ランダム条件については、1 分に 1 度、50% の確率でシステムを起動する判定プログラムを実施するよう設計した。また、ユーザ契機条件のみ事前に被験者に条件を伝え、他の 2 条件は盲検法で実施した。

被験者は、2.2.2 項と同じ 12 名を対象とし、順序効果を相殺するため、各条件の順序を変えた 6 通りの順序で各 2 人ずつ実施する。卓上に実験用 PC および感情センサ [3]、uh-huh duck を配置した環境 (図 2 参照) で、各条件 1 問ずつ、コーディングスキルチェック Web

サイト [4] にてプログラミングタスクを与えた。また実験後の評価のため、各条件間でシステムの印象或使用感、および SUS (System Usability Scale) を調査するアンケートを実施した。

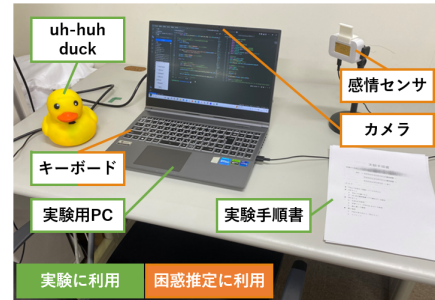


図 2: 実験環境

3.2.2 実験結果

現在までに 12 名中 10 名を対象に実験を終えた。得られたアンケートの結果を図 3 に示す。有意水準 5% でライアン法にて多重比較を行ったところ、Q1 と Q2 に有意な違いが見られた。

また SUS は、ユーザ契機条件で 79.0、システム契機条件で 83.3、ランダム条件で 83.5 となった。SUS の平均値は 68 とされており、全条件で上回ったため、いずれも高いユーザビリティを示すことを確認した。

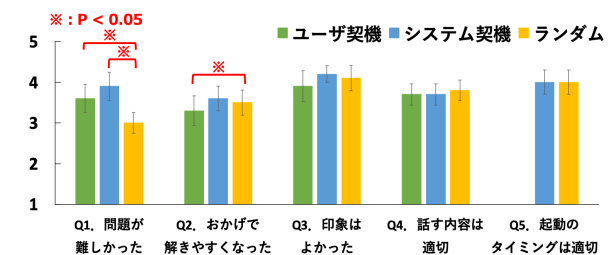


図 3: アンケート結果 (5 段階)

(Q5 はユーザ契機条件では質問せず)

(1: 当てはまらない 3: どちらでもない 5: 当てはまる)

4 考察

アンケートの結果について、Q2、Q3 で、有意差はなかったものの、システム契機条件が他条件を上回っていることが確認できる。要因として、困惑発生を適切に検出して起動できたことが考えられる。ただし、3 条件で問題難易度は同程度に設定したはずが、Q1 の主観的難易度の回答結果には有意差があった。本システムは既報 [1] でも示した通り、主観的難易度が高いほど、有効性が高まるとされており、今回の結果もそれに起因する可能性がある。そのため、困惑検出による起動が直接の要因であったことは断定できない。

5 おわりに

本研究では、困惑状態を推定し、困惑度合に応じてシステム側から使用を促すアヒル人形型システムを開発し、契機の違いによる影響を調査した。実験の結果から、困惑検出による起動が有効である可能性が示された。今後は、条件ごとのシステムの使用回数や、システム使用直後の気付き、およびタスクの成績などにもとづいて、結果の解析を進める。

参考文献

- [1] 倉田 他. 人形との相互作用によるプログラミング時の思考整理支援. HAI シンポジウム 2023.
- [2] 倉田 他. マルチモーダルデータを利用したプログラミング作業中の困惑状態推定. DICOMO 2024 シンポジウム.
- [3] EE Times Japan. 三菱電機、人の感情を推定するセンサーを開発. <https://eetimes.itmedia.co.jp/ee/articles/2209/08/news048.html>. (最終アクセス 2024 年 12 月 30 日)
- [4] paiza. IT エンジニア向け国内最大の転職・就職・学習プラットフォーム. <https://paiza.jp>. (最終アクセス 2024 年 12 月 30 日)