機械学習の適用による社会調査現場での追加情報収集支援システム

高橋 和子　　奥村 学　　鈴木 泰山
敬愛大学国際学部　東工大科学技術創成研究院　（株）ピコラボ
takak@u-keiai.ac.jp　oku@lr.pi.titech.ac.jp　taizan@picolab.jp
清家 大嗣
東大大学院学際情報学府
hirotsugu.seike@koshizuka-lab.org

1　はじめに

社会調査において自由回答で収集されたデータに対し, あらかじめ用意されたコード（カテゴリ）を付与する作業をアフターコーディングと呼ぶ. 代表例として, 社会階層研究で必須の「SSM 職業コーディング」（自由回答を含め多次元で収集される職業情報 [1]に約 200 種類ある職業コード [1] のいずれかを付与）などがある [2].

アフターコーディングは人手（コーダ）により行われるが, 回答に含まれる情報が曖昧であったり不十分な場合は, コーダの負担が増大するだけでなく, 誤ったコードが付与される可能性がある. SSM 職業コーディングにおいては, コーダの労力軽減のためにコンピュータによる自動コーディングシステムが開発され利用されているが [5], ここにも同様の問題が存在する. しかし, コード体系や定義内容を熟知していない回答者や調査員に, どの回答もコードの決定に必要な情報が含まれるよう要求するのは, 現実的とはいえない.

そこで, コンピュータに回答が情報不足であるかどうかを調査現場で判断させ, 不足すると判定した場合は, その場で回答者から有効な情報を追加してもらうシステムの検討を開始した [6]. 現在, カテゴリが用意された自由回答全般への汎用化を考慮しながら, 前述の SSM 職業コーディングを対象としたシステムの構築を進めている [7]. より詳細には, 調査現場に持参したタブレットに回答を入力してシステムが置かれたクラウドサーバに送信し, コーディングに必要な情報が不足するか否かを機械学習を適用した自動コーディングの結果を利用して判定させる. もし不足すると判定された場合は, タブレット画面に候補となる語を提示し, その中から回答者に選択してもらった語を当初の回答に追加する方法である.

提案システムは, 調査現場で回答が電子化されるため, 後日の入力作業が不要となり, またリアルタイムに自動コーディングの結果を得られる [2]という利点ももつ.

本稿の目的は, 提案システムの有効性を, 自動コーディングの結果から実験的に示すことである. 以下, 次節で提案システムの概要を述べ, 3 節で実験結果を報告する. 最後にまとめと今後の課題について述べる.

2　方法（アルゴリズム）

提案システムでは, 回答の情報不足の判定および追加情報の提示方法と収集方法が重要である. アルゴリズムを STEP1〜STEP4 に示す.

STEP1　データ入力とサーバへの送信
STEP2　自動コーディング
STEP3　情報不足の判定（情報不足と判定されなかった場合は STEP4 に進まず終了）
STEP4　追加情報の提示と収集（収集した情報を初期回答に追加し, STEP2 に戻る）

以下で, 各 STEP について説明する.

2.1　データ入力とサーバへの送信（STEP1）

調査現場において, 調査員は回答者から得られた回答をタブレットに入力し, システムを置いたクラウドサーバに送信する.

2.2　自動コーディング（STEP2）

今回は, 既存の SSM 職業コーディングの自動化システム [5] を利用する. すなわち, 機械学習としてサポートベクターマシン（SVM）を one-versus-rest 法により多値分類に拡張した分類器を利用し, 素性として,「従業先の規模を除く職業情報（従業先事業の種類, 仕事の内容, 地位, 役職）, 学歴（選択回答）, SVM を適用する前に実行するルールベース手法により出力されたコー

[1]従業先事業の種類（自由回答）, 仕事の内容（自由回答）, 地位（選択回答）, 役職（選択回答）, 従業先の規模（選択回答）の 5 つを収集する場合が多い.

[2]従来の自動コーディングシステムは入力が完了した後にオフラインでファイル単位で処理を行う.

ド」を用いる．自由回答は，形態素に分解した後，素性番号に変換するが，同じ語でも，出現した場所（従業先事業の種類か仕事の内容か）で素性番号を変える．

自動コーディングシステムでは，第 1 位に予測された結果に対し，信頼度の目安となる「確信度」を 5 段階で付与する [5]．各確信度は，レベル A の信頼度が最も高く，レベル E が最も低くなるように，[4] のアイディアに従って，複数の分類スコアを利用して決定する．ここで，rank1, rank2 は，それぞれ SVM により第 1 位，第 2 位に予測されたコードに付随して出力される分類スコア（分離平面からの距離）を示す．また，α, β は閾値で，2005 年 SSM 調査データセット（12,500 事例）を用いた実験により，$\alpha = 3$, $\beta = 0.4$ とした．

A:$rank1 >= 0$ かつ $rank2 < 0$, $rank1 - rank2 >= \alpha$

B：$rank1 >= 0$ かつ $rank2 < 0$, $rank1 - rank2 < \alpha$

C：$rank1 >= 0$ かつ $rank2 >= 0$

D：$rank1 < 0$ かつ $rank2 < 0$, $rank1 - rank2 >= \beta$

E：$rank1 < 0$ かつ $rank2 < 0$, $rank1 - rank2 < \beta$

現時点で自動コーディングシステムが対象とする 4 種類の職業 · 産業コーディング [3]における確信度ごとの正解率（全事例中，正解であった事例の占める割合）と出現率（全事例の中で 該当する事例が出現した割合）は表 1 の通りであった [4] [6] ．表 1 より，確信度ごとの正解率は，コーディングの種類やコードの数が異なっていても（該当する事例が出現しなかった ISIC のレベル C を除く），レベル $A >$ レベル $B >$ レベル $C >$ レベル $D >$ レベル E の順で安定していることがわかる．

2.3 情報不足の判定（STEP3）

調査現場でコードを，決定するための情報が回答に含まれているか否かを調査員が確認することは現実的ではない．そこで，以下に述べるように，確信度を利用した自動判定を行う方法を提案する．

SSM 職業コーディングの自動化システムは，ルールベース手法により予測されたコードを SVM の素性として用いるが [3]，[6] によると，ルールベース手法によるコード決定率は，SVM の結果から算出される確信度のレベルが下がるほど低かった [5]．コーダは，基本的に，

[3]ISCO（International Standard Classification of Occupation）と ISIC（International Standard Industiral Classification of All Economic Activities）は ILO が定めた国際標準職業分類と国際標準産業分類である．

[4]評価事例はいずれも JGSS-2006 データセット（計 2,203 事例）であるが，訓練事例は，SSM 職業 · 産業コーディングでは JGSS-2000, -2001, -2003, -2005 データセット（27,585 事例），ISCO·ISIC コーディングでは 2005SSM 調査データセット（16,089 事例）を用いた．この理由は，ISIC コードは 2005SSM 調査と JGSS-2006 以外の調査では付与されていないためである．

[5]例えば，表 1 において，ルールベース手法によるコード決定率は，全体では 83.2% であったが，確信度ごとにみると，レベル A の 99.7% から確信度レベル E の 32.5% まで順に低下した．

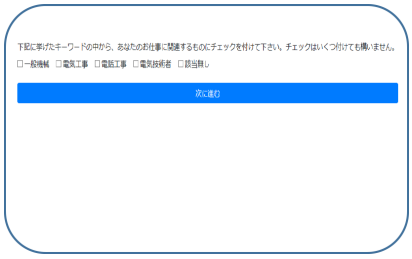

図 1: 追加情報収集用画面例

ルールベース手法が根拠とするマニュアル [1] に記載されたルールに基づいて判断するため，ルールベース手法をコーダの代理とみなすことにすると，確信度レベルが低い回答ほどコーダの判断も困難であったと考えられる．他方で，コーダの判断が困難な原因は，回答の情報不足という問題が大きいと考えられるため，回答が情報不足か否かの判定には，確信度の利用が有効であると判断した．

確信度レベルを利用する際に，どのレベルまでとするかについては，表 1 からも明らかなように，確信度レベルがもっとも低い E はどのような場合でも正解率が安定して非常に低いのに対し，レベル D はバラツキが大きいため，確信度レベル E に限定する．

以上より，回答の情報不足は，「確信度がもっとも低いレベル（今回はレベル E)」の場合とする．

2.4 追加情報の提示と収集（STEP4）

実際には，調査現場では，得られたコードが正解か否かは不明であるが，提案システムでは，確信度レベルが E の場合は不正解とみなし，得られたコードに依存した選択肢（語レベル）を提示し，回答者に選んでもらう（図 1 参照）．追加する情報を自由回答で収集しない理由は，回答空間を大きくしないためである．

選択肢は，まず過去の事例からあらかじめ作成した「不正解－正解のペア表」を検索して対応する正解コードを見つけ，次にこの正解コードにより，あらかじめ作成した「正解対応カテゴリ名表」を検索して見つけたカテゴリ名とする方法がある（方法 1）．例えば，図 1 は，入力された回答 [6]に対して予測されたコードが「628」（鋳物工，鍛造工，金属材料製造作業者）であったため，不正解－正解のペア表により，正解コードとして「633」（一般機械器具組立工 · 修理工）などを見つけ，そのカテゴリ名を提示した画面である．

[6]回答は，「従業先事業の種類（自由回答）：部品の製造，仕事の内容（自由回答）：工場で部品の検査，地位：わからない，役職：わからない，従業先の規模：1,000 人〜1,999 人，学歴：新制高校」である．

表 1: コーディングの種類別確信度ごとの正解率（カッコ内は出現率）

コーディングの種類	コード数	確信度 A	確信度 B	確信度 C	確信度 D	確信度 E
SSM 職業（小分類）	約 200	0.954(0.29)	0.716(0.48)	0.505(0.09)	0.402(0.04)	0.226(0.10)
SSM 産業（大分類）	約 20	0.975(0.32)	0.867(0.54)	0.646(0.02)	0.603(0.06)	0.336(0.06)
ISCO（小分類）	約 400	0.963(0.05)	0.701(0.67)	0.444(0.06)	0.336(0.05)	0.201(0.17)
ISIC（亜大分類）	約 60	0.941(0.01)	0.919(0.56)	- (0.00)	0.812(024)	0.240(0.19)

方法 1 はカテゴリ名がある場合に限定されるが, 提示する選択肢を決める方法は, 以下に示すように他にもある（例はすべて SSM 職業コーディングの場合）.

方法 1：カテゴリ名

（例）SSM 職業小分類名　自然科学研究者 [7]

方法 2：混同されやすいコード間で決め手となる語 [8]

（例）「557」と「573」の場合　内勤, 外回り

方法 3：コードを特徴づける語

（例）「501」の場合　研究員, 研究所, 試験研究

方法 4：不正解コードと正解コードを弁別する語

（例）不正解「501」正解「688」の場合　雑用, 作業

方法 2 は, 過去の経験から得られた知識に基づくが, 用意されたコード対にしか適用できない. 方法 3 と方法 4 は, 自動コーディングシステムで用いる訓練事例（約 50,000 事例）を利用し, tf-idf 法により機械的に語を抽出する. それぞれ方法 1 と方法 2 と類似する.

この他, 予測コードに関係なくつねに職業大分類のカテゴリ名 [8]を選択肢とする非常に簡単な方法もある [6].

選択肢は複合語で提示する方が適切な場合も多いが, 初期回答に追加する際は, 形態素に分解後, 素性番号に変換する. STEP4 の課題は, 提示すべき情報が多過ぎる場合の対応である.

3　実験

提案システムにおいて, 情報不足の判定に確信度レベル E を用いることについては, [6] により実験的に有効性が示されている.

本稿では, 追加情報の提示と収集方法として, 2.4 節で提案した 4 つの方法を初期回答や大分類語を追加する方法 [6] と比較することで有効性を示す. ただし, 今回の実験は, 回答者が選んだ語ではなく, すでに付与された正解から回答者が選ぶ語を推測したため, 実際より正解率が高い結果となる可能性がある.

[7]「501」のカテゴリ名であるが, この後に「502」人文科学研究者, · · ·, 「688」その他の労務作業者と続く. 1 つのコードに複数の名称が対応するものもある

[8]専門的 · 技術的職業, 管理的職業, 事務的職業, 販売的職業など計 15 個ある.

3.1　実験設定

実験では, 訓練事例は, JGSS-2000, -2001, -2003, -2005, -2006 データセット（計 33,711 事例）, 評価事例は, 訓練事例と性質が同じ JGSS-2008 データセット（2,662 事例）と性質が異なる JLPS 第 1 波 [9]データセット（4,345 事例）のうち, 確信度レベルが E と判定された 274 事例（全体の 10.3%）と 723 事例（全体の 16.6%）の 2 つを用いる. 正解率は, それぞれ 26.6% と 28.2% である. 評価尺度は, 正解率と確信度を用いる.

3.2　実験結果

最初に, 確信度レベル E が付与された全事例の正解率を図 2 に示す. 方法 3 で特徴語が取り出せたコードは 163 個（82.7%）であり特徴語が追加できない事例は, JGSS-2008 データセットで 2.6%, JLPS 第 1 波データセットで 1.9% 存在した. 正解率の算出はいずれも全事例を分母としたため, 方法 3 は他の方法より不利な条件であるが, 有効性がもっとも高く, 初期回答より 16%〜28% ポイント向上した. これより, 実際の回答から生成された訓練事例に出現した語を利用する方が, 抽象度の高いカテゴリ名を利用するより有効であると考えられる. 確信度は, 初期回答ではすべて E であったが, 方法 1, 3 とも約 7 割の事例がレベル D 以上に向上した.

正解率をさらに改善するには, 各方法をうまく組み合わせることが有効になると考えられるが, このためには, 各方法の関係について調査しておく必要がある. 今回, 方法 1 と方法 3 の正解 · 不正解の状況を調査した結果, 表 2 に示すように, 約 8 割の事例で一致した. 両者の相違点についての詳細な調査は今後の課題である.

次に, 方法 1 と方法 2 における正解率を図 3 に示す. 方法 2 には「557」と「573」, 「607」と「686」, 「633」と「634」「635」, 「601」と「631」が存在し, JGSS-2008 データセットで 51 事例（18.6%）, JLPS 第 1 波で 158 事例（21.9%）が該当した. 図 3 より, 方法 2 の有効性は示せたが, 該当事例が少ないため, 単独で用いるのは効果的ではない.

[9]東京大学社会科学研究所が 2007 年に実施した「働き方とライフスタイルの変化に関する全国調査」の若年・壮年パネル調査である.

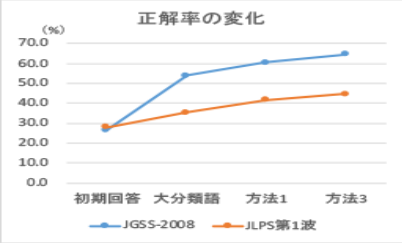

図 2: 正解率（初期回答, 大分類語追加, 方法 1, 方法 3）

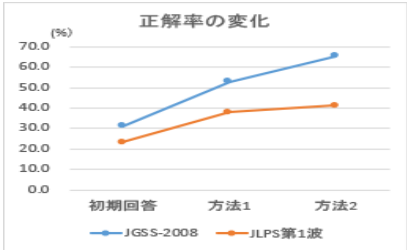

図 3: 正解率（初期回答, 方法 1, 方法 2）

表 2: 方法 1 と方法 3 の正解 · 不正解状況

		［方法 3］正解	不正解	計
	正解	53.6%	6.9%	60.6%
［方法 1］	不正解	10.9%	28.5%	39.4%
	計	64.6%	35.4%	100.0%

最後に, 方法 1 と方法 4 の正解率を図 4 に示す. 方法 4 で, あらかじめ生成できた正解－不正解ペアとマッチしたのは, JGSS-2008 データセットで 33 事例（12.0%）, JLPS 第 1 波で 105 事例（14.5%）しかなかった. 方法 4 の有効性は示せたが, 方法 2 と同様に単独での利用は効果的ではない.

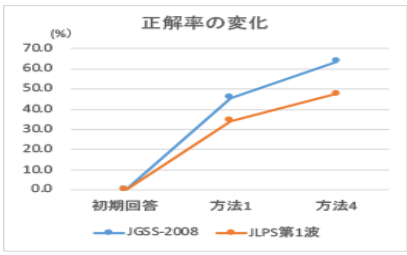

図 4: 正解率（初期回答, 方法 1, 方法 4）

4　おわりに

本稿では. 社会調査において収集される自由回答に, 分類に必要な情報が含まれているか否かを機械学習により調査現場で判断し, 不足すると判定した場合は, その場で回答者から有効な情報を追加してもらうシステムを提案し, 実験的に有効性を示した.

今後の課題は. システムの精度を向上させ, 実際の利用者であるコーダや調査員に評価してもらうことである. このためには. 実装に向けて追加情報の提示方法を工夫する必要がある.

謝辞　2005 年 SSM 調査データの利用に関して, 2005 年 SSM 調査研究会の許可を得た. 日本版 General Social Surveys（JGSS）は, 大阪商業大学 JGSS 研究センター（文部科学大臣認定日本版総合的社会調査共同研究拠点）が, 東京大学社会科学研究所の協力を受けて実施している研究プロジェクトである. 東大社研パネル調査プロジェクトにおける職業・産業コーディングの精度向上を目的として, 職業・産業の自由記述データの提供を受けた. 本研究は JSPS 科研費（16k04039）の成果の一部である.

参考文献

[1] 1995 年 SSM 調査研究会. 2006. SSM 産業分類 · 産業分類（95 年版）.

[2] 原純輔. 1984. 社会調査演習. 東京大学出版会.

[3] 高橋和子, 高村大也, 奥村学. 2005. 機械学習とルールベース手法の組み合わせによる自動職業コーディング. 自然言語処理, 12(2), pp.3–24.

[4] K. Takahashi, H. Takamura, and M. Okumura. 2008. Direct estimation of class membership probabilities for multiclass classification using multiple scores. In *Knowl Inf Syst* 19(2), pp.185–210. Springer London.

[5] 高橋和子, 多喜弘文, 田辺俊介, 李偉. 2017. 社会学における職業 · 産業コーディング自動化システムの活用. 自然言語処理, 24(1), pp.135–170.

[6] 高橋和子. 2018. 機械学習を適用した自由回答収集時における有効情報追加システムの構想－職業コーディングを例として－. データ分析の理論と応用, 7(1), pp.21–42.

[7] 高橋和子, 奥村学. 2019. 機械学習の適用による調査現場での追加情報収集支援システムの構築. 数理社会学会第 68 回年次大会報告要旨, pp.38–41.

[8] 豊田真之, 常松淳. 2008. コーディング作業における留意点. 東大社会科学研究所パネル調査プロジェクト ディスカッションペーパーシリーズ NO.6 職業 · 産業コーディングマニュアルと作業記録, pp.53–58.