2F-04

機械学習を適用した自由回答のコーディング支援
-職業・産業コーディング自動化システムとその拡張-

高橋　和子　　多喜　弘文　　田辺　俊介　　李　偉
敬愛大学国際学部　法政大学社会学部　早稲田大学文学学術院　東工大理工学研究科

１．はじめに

社会調査における回答形式には，提示された選択肢の中から選ぶ選択回答と，自由に記述する自由回答があるが，自由回答は統計処理を行う際にコードに変換する手間を要するため，可能な限り選択回答を用いることが推奨される[1]．ただし，国勢調査でも行われているように，職業や産業情報は例外で，自由回答で収集したものを研究者自身がコードに変換する場合が多い．この作業は「職業・産業コーディング」と呼ばれるが，分類コードが多く（例えば職業コードの場合、国内標準コード約 200 個，国際標準コード約 400 個），コーディングのルールも複雑なために，特に大規模調査の場合はコーダの作業負担が膨大となる．また，コーダが熟練していない場合は，コーディングの結果に一貫性がない可能性が高いことも指摘されている．

この問題を解決するために，機械学習に自然言語処理に基づいたルールベース手法を組み込んだコーディング自動化システムを開発した[2]．本システムの利用により，コーダはサポートベクターマシン（SVM）により予測されたコードを参考にしながらコーディングが行える．その後，第 1 位の予測コードに対しては確信度を付与する機能も追加した．現在，本システムは東大社会科学研究所附属社会調査・データアーカイブ研究センター（SSJDA）の Web を通じて試行提供されている（http://csrda.iss.u-tokyo.ac.jp/joint/autocode/）[3][4]．本稿では，本システムについて利用方法も含めて報告した後，Web 公開後に追加したメンテナンスの自動化機能についても述べる．さらに，本システムを他の自由回答に拡張する方法についても検討する．

２．関連研究

職業・産業コーディングにおける前述の問題は海外においても認識されており，近年，韓国や米国でも自動化システムが開発された．いずれも自動化アルゴリズムが単純で，例えば大韓民国統計庁の a Web-based AIOCS(Automated System for Industry and Occupation Coding)は，ルールベース手法，最大エントロピー法，情報検索技術[5]，米国 CDC（Centers for Disease Control and Prevention）の SOIC はデータベースとの単語のマッチング[6]，SOIC の後継である NIOCCS(NIOSH Industry and Occupation Computerized Coding System)はルールベース手法[7]の適用にとどまる．

３．確信度付き職業・産業コーディング自動化システムと利用方法

3.1 変換可能な職業・産業コードと自動化の手法

本システムで変換できるコードは表 1 最左列に示す 4 種類で，利用者は最大 4 種類まで自由に選べる．SSM 職業・産業コードは社会調査における国内標準コード，ISCO（International Standard Classification of Occupations），ISIC（International Standard Industrial Classification of All Economic Activities）は ILO により定められた国際標準コードである．本稿では，従業先事業内容（自由回答），仕事内容（自由回答），地位・役職（選択回答）をまとめて「基本素性」とよぶ．

表 1　変換可能なコードと自動化の手法

コードの種類	コードの数	自動化の手法（SVM において用いる素性）
SSM 職業コード（小分類）	約200	ルールベース手法とSVM の組み合わせ（**基本素性**，ルールベース手法の結果）
SSM 産業コード（大分類）	約20	ルールベース手法とSVM の組み合わせ（**基本素性**，ルールベース手法の結果）
ISCO（小分類）	約 400 階層構造（4 層）	SVM（**基本素性**，学歴，SVM により第1 位に予測されたSSM 職業コード）
ISIC（亜大分類）	約 60 階層構造（4 層）	SVM（**基本素性**，SVM により第1 位に予測されたSSM 産業コード）

すべてのコードが希望された場合の処理を示す．い

A Coding Support System for Open-ended Survey Data with Machine Learning: An Automatic Occupation and Industry Coding System and its Extensibility
Kazuko TAKAHASHI (takak@u-keiai.ac.jp)
Department of International Studies, KEIAI University
Hirofumi TAKI (taki@hosei.ac.jp)
Faculty of Social Sciences, HOSEI University
Shunsuke TANABE (tanabe.sh@waseda.jp)
Faculty of Letters, Arts and Sciences, WASEDA University
Wei LI (li.w.aa@m.titech.ac.jp)
Graduate School of Science and Engineering, Tokyo Institute of Technology

ずれも第1位から第3位までの予測コードを決定する.

STEP1　職業・産業情報に対する形態素解析[8]

STEP2　ルールベース手法により，仮SSM職業コードと仮SSM産業コードを出力

STEP3　基本素性とSTEP 2で出力された仮SSM職業コードを素性としてSVMを適用し，SSM職業コードを決定

STEP4　基本素性，学歴，STEP3で決定されたSSM職業コード（第1位のみ）を素性としてSVMを適用し，ISCOを決定

STEP5　基本素性とSTEP2で決定された仮SSM産業コードを素性としてSVMを適用し，SSM産業コードを決定

STEP6　基本素性とSTEP5で決定されたSSM産業コード（第1位のみ）を素性としてSVMを適用し，ISICを決定

3.2 確信度の付与機能

確信度は，「A：人手によるコーディングは不要 B：できれば人手によるコーディングを行う方がよい C：人手によるコーディングが必要」の3段階とし，決定条件は次の通りとした．ただし，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伴って出力されるスコア（分離平面からの距離）を示す．αは閾値で，今回はα=3とした．実験の結果，確信度別の精度は順に約98%，76%，40%程度であった．

A：rank1 > 0 かつ rank2 <= 0，rank1 − rank2 > α

B：rank1 > 0 かつ rank2 <= 0，rank1 − rank2 <= α

C：A，B以外の場合

3.3 システムの利用方法

利用者は，「通し番号，学歴，従業上の地位・役職，従業先事業内容，仕事内容，従業先規模」の順にデータを並べた入力用データファイルを準備し，SSJDAを通じた以下の手続きにより自動コーディング結果を得る．

(1)［利用者］利用申請書をメイルによりSSJDAに送信（希望する職業・産業コードの種類を明記）

(2)［SSJDA］ユーザID，パスワードの発行およびファイルのアップロード／ダウンロード場所の通知

(3)［利用者］入力用データファイル（CSV形式）を指定場所にアップロード

(4)［SSJDA］図1右下Runボタンにより本システムを実行し，結果ファイル（CSV形式）を生成

(5)［利用者］指定場所から結果ファイルをダウンロード

4．メンテナンス自動化機能ー訓練事例の更新ー

職業・産業コーディングが実施されるたびに正解（最終的に決定されたコード）付きの事例が蓄積される．これらを既存の訓練事例に追加すれば，訓練事例の量が増える上に新しい事例への対応も高まるため，精度向上が期待できる．システムの継続性を考慮すると，この作業を誰もが容易に行える必要があると考え，以下に示す訓練事例更新の自動化機能を追加した．

(1) 入力用データファイルの最右列に正解を入力した（複数列に複数種類のコードも可）「正解付きファイル」（CSV形式）を準備

(2) 正解付きファイルを図1右上Openボタンで指定し，訓練事例の該当コードをチェック

(3) 図1左下Updateボタンを押す

図1　SSJDA側操作用画面

5．システムの拡張

本システムを，自由回答（最大2種類）と選択回答（最大3種類）が混在または自由回答のみのデータを対象とするシステムに拡張することは比較的容易で，現在開発中である．ただし，実際の適用では，機械学習のための正解付き訓練事例を用意する必要がある．

6．おわりに

本稿では，SSJDAのWebを通じて利用できる職業・産業コーディング自動化システムと，メンテナンスの自動化機能について述べた．今後の課題は，本システムにおける精度向上を目指しつつ，より一般的な自由回答のコーディング支援システムに拡張することである．

謝辞

2005年SSM調査データの利用に関して，2015年SSM調査研究会の許可を得た．本研究はJSPS科研費25380640の助成を受けたものである．

参考文献

[1] 原純輔・海野道郎，1984，社会調査演習，東京大学出版会.

[2] Takahashi, K. et al., 2005, Automatic occupation coding with combination of machine learning and hand-crafted rules. *LNAI* Vol.3518, 269-279. Springer Berlin, Heidelberg.

[3] Takahashi, K. et al., 2014, An Automatic Coding System with a Three-Grade Confidence Level Corresponding to the National/International Occupation and Industry Standard: Open to the Public on the Web, In *Proceedings of the Sixth International Conference on KEOD 2014*, 369-375.

[4] 高橋他，2014，社会調査における職業・産業コーディング自動化システムの一般公開と運用，言語処理学会第20回年次大会論文集，932-935.

[5] Jung, Y. et al., 2008, A web-based automated system for industry and occupation coding, In *Proceedings of the Ninth International Conference on WISE 2008*, 443-457.

[6] SOIC, 2000, http://www.cdc.gov/niosh/soic/default.html (accessed December 25, 2015)

[7] NIOCCS, 2013, http://wwwn.cdc.gov/niosh-nioccs/ (accessed December 25, 2015)

[8] 黒橋禎夫・長尾真，1998，日本語形態素解析システムJUMAN version 3.61，京都大学大学院情報学研究科