Web版職業・産業コーディング自動化システムの開発

高橋 和子　田辺 俊介　吉田 崇
敬愛大学国際学部　東京大学社会科学研究所　静岡大学人文社会科学部
takak@u-keiai.ac.jp　tanabe@iss.u-tokyo.ac.jp　jtyoshi@ipc.shizuoka.ac.jp
魏 大比　李 偉
名校教育グループ　東工大大学院理工学研究科
dabiwei@s-crams.com　li.w.aa@m.titech.ac.jp

1　はじめに

社会調査においては, 個人の仕事内容である「職業」や, 従業先の事業内容である「産業」は重要な情報であり, 正確性を期するために自由回答で収集される場合がある. このとき, 統計処理を行うためには, 自由回答をあらかじめ決められた分類コードに変換する作業（職業・産業コーディング）[1]が必須であるが, 職業や産業のコードは個数が多く[2], コード化のルールも複雑なために[3], コーダにとって多大な労力や時間を要するという問題が存在する [4].

そこで, コーダを支援する目的で, 職業・産業コーディングにルールベース手法や機械学習であるサポートベクタマシン（SVM）を適用して自動化を行い, その結果を候補として提示するシステムが開発された [7, 8, 9]. システムはSSM（Social Stratification and social Mobility）調査やJGSS（Japanese General Social Surveys）などの大規模調査を中心に利用されているが, 公開されていないために, 研究者個人では利用しにくい状況にある.

他方で, 近年の国際比較研究の隆盛に伴い, 前述した国内標準コード以外に国際標準のコードも必要になってきた. このため, 新たな自動化システムの開発が試みられたが [10], 前述のシステムとは独立に存在するため, 国内・国際標準コードが両方必要な場合は, 2つのシステムを別々に稼働させなければならない.

さらに, システムの開発当初は, コーダは自動コーディング結果を参考にしながらも, すべての事例についてコーディングを行うことが想定されていたが, 最近は, 自動コーディング結果の信頼性が高い事例に対してはこれを省略することでコーダの作業量を軽減したい場合があり, この対応が要請されている[4].

これらの問題のうち, 最初の2つについては, これまで開発された種々の自動化システムを整理・統合し, 利用しやすい形でWeb公開を行うことで解決できる. また, 最後の問題であるコーダの作業量軽減のためには, 人手によるコーディングの必要性を判断する基準として, 自動コーディングの結果に確信度を示す値を付与することが有効であると考えられる.

以上を踏まえ, 利用者が, 東京大学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サイト[5]を通じて, 職業や産業情報を所定の形式のデータファイルとしてアップロードすれば, 希望する職業・産業コードの自動コーディング結果と結果に対する確信度が付与されたファイルをダウンロードできるシステムを開発中である. 本稿では, このシステムについて述べる.

以下, 次節で, 我が国の社会調査において利用されている職業・産業のコード体系について述べ, 3節で関連研究について述べる. 4節で本システムについての説明を行い, 最後にまとめと今後の課題について述べる.

2　職業・産業のコード体系

各国において用いられる職業・産業のコード体系はさまざまであるが, 我が国の社会調査では, 国内標準コードとして, 国勢調査で用いられる日本職業標準分類・産業分類に基づいて独自に作成されたSSM職業小分類（以下, SSM職業コードとよぶ）およびSSM産業大分類（以下, SSM産業コードとよぶ）[1] が用いられる（表1参照）. ここで, 職業の方が産業より分類レベルが細かい理由は, 社会学においては, 個人の情報を示す職業の方が産業より重要な変数であるためである.

国際標準コードとしては, 国際労働機構により作成されたISCO（International Standard Classification of Occupation）およびISIC（International Standard Industrial Classification of All Economic Activities）[3]

[1]たとえば, 仕事の内容が「市役所庶務課で事務員（内勤）」と回答された場合, 職業コード「５５４」（総務・企画事務員）にコーディングする.

[2]職業は数百個, 産業は数十個のコードに分類されることが多い.

[3]職業は, 役職や従業上の地位, 従業先事業の規模など仕事の内容以外の情報も用いられて総合的に判断されるため [4], たとえば前述の例において, 役職が「係長、係長相当職」（選択肢）であれば「５５４」であるが, 「課長、課長相当職」（選択肢）であれば「５４５」（管理的公務員）になる [2].

[4]このような要請は, 国勢調査のように事例が膨大に存在する場合だけでなく, 事例がそれほど多くなくても, 熟練したコーダがいなかったり, 時間的コストがかけられないような場合にも生じる.

[5]http://ssjda.iss.u-tokyo.ac.jp/

表 1: 我が国の社会調査で用いられる職業 · 産業コード

	職業 · 産業コード	分類レベル	個数
国内標準	SSM 職業コード	小分類	約 200
	SSM 産業コード	大分類	20
国際標準	ISCO	小分類	約 400
	ISIC	亜大分類	60

の 88 年版が用いられる. 国際標準コード体系が国内標準コード体系と大きく異なるのは, 階層的な構造であることである. また, ISCO の決定に, 教育レベルを判断基準とする「スキルレベル」が設定されている点も異なる. 我が国の社会調査では, ISCO は小分類（4 桁）, ISIC は亜大分類（2 桁）までを必要とすることが多い.

3 関連研究

ここでは, これまでに国内で開発された職業 · 産業コーディング自動化システムおよび海外（韓国, 米国）における Web 公開システムの状況について述べる.

まず, 国内で開発された種々のシステムについては, Web 公開を行うものに限定して述べる. 最初の自動化システムは, 職業や産業情報 [6]に格フレームの考え方を適用したルールベース手法によるシステム（ROCCO システム）である [7]. これにより, SSM 産業コードの正解率は約 75% であったため, 現在も SSM 産業コードには ROCCO システムが適用されている. ここで, 本稿における正解とは, 人手による職業 · 産業コーディングの実施により最終的に決定されたコードをいう. 職業 · 産業コーディングでは全事例に正解が付けられるため, 本稿では再現率を正解率とよぶ.

ROCCO システムでは, SSM 職業コードの正解率は 70% に満たなかったため, 職業 · 産業情報のうち, 仕事の内容, 従業先の事業内容, 従業上の地位, 役職（以後, 基本素性とよぶ）に, ROCCO システムの結果を素性として追加して SVM を適用するシステム [8] を開発した [7]. この結果, 正解率は 80% [8]に向上したため, SSM 職業コードには [8] のシステムが適用されている.

ISCO は SSM 職業小分類と単純な変換関係が存在しないため [13], 新たな自動化システムを開発する必要があった. 基本素性のみを素性とする SVM によるシステムの正解率は 60% に満たなかったが, スキルレベルに関連が深いと考えられる学歴と, [8] のシステムにより出力される SSM 職業コード 3 個を素性に追加した結果, 第 3 位に予測されたコードまでを含めると 70% に向上した [10]. ISCO の正解率が低いのは, コード数が多い上に, 正解付きの事例が少なかったためであるが, 今後, ISCO コーディングの普及による向上が見込める. ISIC については, 有効な素性選択について実験中である.

次に, 大韓民国統計庁において Web 公開が検討されている産業 · 職業情報の自動化システム（Web-based AIOCS）[5] について述べる. 韓国における職業 · 産業コードは大韓民国統計庁により作成され, いずれもレベル 4 までの階層構造をもつ. 職業コードと産業コードの数は, それぞれ 442 個, 450 個である. ルールベース手法, 最大エントロピー法（MEM）, 情報検索技術（IRT）の 3 種類を用意し, 単独またはルールベース手法と他の 2 つの方法のいずれかまたは両方を組み合わせた計 6 種類の方法が存在する. ただし, 組み合わせの意味が本システムとは異なり, 手法自体は独立のままで, ルールベース手法によりルールがマッチしなかった場合に別の手法を実行する. 単独の方法より複数の方法を組み合わせた方が性能がよく, 特に, ルールベース手法, MEM, IRT を順に実行する方法の精度は 98.4% である. しかし, コードの決定率が高くないために, 正解率は 76.3% である.

ユーザインターフェイスは, anonymous user も対象とするためか, 一問一答の伝票形式画面で, 会社名, ビジネスカテゴリ, 部門, 役職, 仕事の内容（自由回答）を入力すると, 同一画面に結果が表示される. これに対し本システムでは研究者を想定し, ファイルによる入出力を行う.

米国においては, CDC（Centers for Disease Control and Prevention）により, SOIC（Standardized Occupation & Industry Coding）システムが Web で公開されており, ソフトウェアをダウンロードできる [9]. ルールベース手法によるマッチングが主で, 正解率は職業コード 75%, 産業コード 76% で, 両方では 63% である.

4 Web 版職業 · 産業コーディング自動化システム

4.1 システムの構成と処理手順

本システムの構成図を図 1 に示す. 各自動化システムは表 1 に示した各コードに対応しており, 利用者の希望に応じて複数のコードを一度に提供することができる.

[6]仕事の内容 (自由回答）, 従業先の事業内容 (自由回答）, 従業上の地位（選択回答）, 役職（選択回答）, 従業先事業の規模（選択回答）から構成される.

[7]SSM 職業コードは多値分類であるため, SVM は one-versus-rest 法により多値分類器に拡張した.

[8]この値は, 熟練したコーダには及ばないものの一般のコーダより高い.

[9]http://www.cdc.gov/niosh/soic/SOIC.About.html

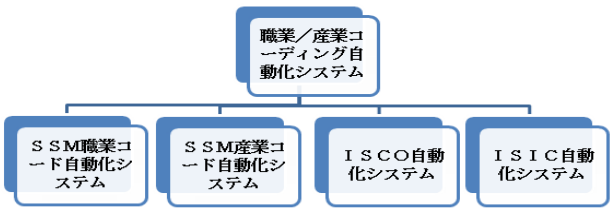

図 1: システムの構成図

本システムの処理手順を STEP 1～STEP 5 に示す. ここで, SVM は, 関連研究で述べたシステムと同様に, one-versus-rest 法により多値分類器に拡張した.

STEP 1　職業・産業情報に対する形態素解析 [6].

STEP 2　ROCCO システムの適用により, SSM 職業コードと SSM 産業コードを出力. SSM 産業コードを決定.

STEP 3　基本素性に STEP 2 により出力された SSM 職業コードを追加して SVM を適用し, SSM 職業コードを決定.

STEP 4　基本素性に学歴と STEP 3 により決定された SSM 職業コード（第 1 位のみ）を追加して SVM を適用し, ISCO を決定.

STEP 5　STEP 2 により決定された SSM 産業コードを素性に追加して SVM を適用し, ISIC を決定（未完成）.

STEP 3～STEP 5 において決定されるコード（SSM 職業コード, ISCO, ISIC）には, システムの確信度が付与される. 次節で説明する.

4.2　確信度の付与

本システムでは, 確信度を 「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 種類に区別する. 各確信度の決定条件は次の通りである [10]. 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ともなって出力されるスコア（分離平面からの距離）を示す. また, α は閾値であり, 本稿では $\alpha = 3$ とした.

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

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

C：A, B 以外の場合

[10] [11] により SVM により予測されたクラスに対するクラス所属確率を推定する方法が提案されているが, 手続きが煩雑なため, 本システムでは基本的な考え方を踏襲し, より簡便な方法を提案する.

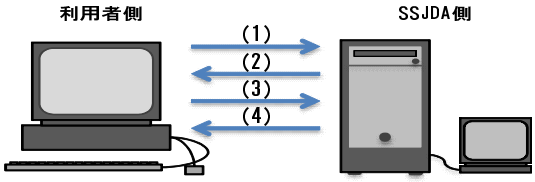

図 2: 利用手順

図 3: SSJDA 担当者操作用画面

4.3　利用方法と本システムの位置づけ

利用者は, 所定の形式の入力用データファイル [11] を用意し, （1）～（4）の手続きにしたがって自動コーディングの結果を得る（図 2 参照）. 本システムは（3）と（4）の間に位置し, SSJDA の担当者が稼働させる. 図 3 は, 本システムを稼働させたときの初期画面を示す.

（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をアップロード

（4）［利用者］結果ファイル（CSV 形式）をダウンロード

4.4　システムの評価

本システム開発の目的は, 利便性の向上およびコーダの作業量軽減であるが, ここでは正解率と確信度付与の有効性について報告する. 評価は現実の場面を想定し, SVM における訓練事例には過去の事例を用いた. すなわち, 評価事例は JGSS06 調査（2,206 サンプル）, JGSS08 調査（1,357 サンプル）, JGSS10 調査（2,579 サンプル）, 2005SSM 調査（16,089 サンプル）の各データセットを用い, 訓練事例は, SSM 職業コード用には JGSS00, 01, 02, 03, 05 調査を合計したデータセット（39,120 サンプル）, ISCO 用には 2005SSM 調査のデータセットを用いた. SSM 産業コードはルールベース手法のみを適用したため, 正解率のみを示す.

正解率を表 2 に示す. コード別の平均は, SSM 職業コード 0.792, SSM 産業コード 0.731, ISCO0.711 であっ

[11] ID, 学歴, 従業上の地位と役職, 仕事の内容, 従業先の事業内容. 従業先の規模の順にデータが入力された CSV 形式のファイルである.

表 2: 正解率（第 3 位に予測されたコードまで含む）

コード	JGSS06	JGSS08	JGSS10	SSM
SSM 職業	0.788	0.789	0.783	0.806
SSM 産業	0.709	0.776	0.739	0.701
ISCO	0.722	0.721	0.691	-

表 3: 確信度別の正解率（カッコ内はカバー率）

コード	A	B	C
SSM 職業	0.954(0.29)	0.716(0.48)	0.355(0.23)
ISCO	0.940(0.07)	0.677(0.67)	0.284(0.26)

た. SSM 職業コードを 3 節で述べた [8] と比較すると, 第 3 位の予測コードまでを含み, 訓練事例のサイズが増大したにもかかわらず, 約 1% 低下した. 素性の一つである ROCCO システムの正解率も約 5% 低下したため, 今後, シソーラスやルール辞書の更新を行う必要がある.

4.2 節で述べた条件により決定した確信度別の正解率とカバー率を表 3 に示す. カバー率は確信度が付与されたサンプルが全サンプルに占める割合である. 表中の数値は, 表 2 に示した評価事例における平均値である. 確信度 A が付与された場合のカバー率が大きいほどコーダの作業が軽減できるが, 決定条件を変化させた実験の結果, 正解率とトレードオフの関係があった.

5 おわりに

本稿では, 多大な労力と時間を要する職業 · 産業コーディングにおいて, コーダを支援するために, 確信度が付与された国内 · 国際標準コードを一度に提供する Web 公開版自動化システムについて述べた.

今後の課題は, まず ISIC 自動化システムを本システムに組み込むこと, 次に, ROCCO システムの改善により, SSM 職業コード自動化システムの正解率を向上させることである. また, ISCO 自動コーディングについて, すでに SSM 職業コードが決定ずみの既存の調査データに対する要請が生じてきたため, この機能を追加したい.

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. 本研究は科研費（22530516）の助成を受けたものである.

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] 1995 年 SSM 調査研究会. 2006. 1995 年 SSM 調査コード · ブック.

[3] Bureau of Statistics; International Labour Office. 2001. Coding Occupation and Industry. Bureau of Statistics; International Labour Office.

[4] 原純輔. 1984. 社会調査演習. 東京大学出版会.

[5] Y. Jung, J. Yoo, S-H. Myaeng and D-C. Han. 2008. A Web-based Automated System for Industry and Occupation Coding. In *Proceedings of the Ninth International Conference on Web Information Systems Engineering (WISE-08)*, LNCS, pp.443–457.

[6] 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.

[7] 高橋和子. 2000. 自由回答のコーディング支援について －格フレームによる SSM 職業コーディング自動化システム－. 理論と方法　Vol.15 No.1, pp. 149–164.

[8] 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　Vol.12 No.2, pp. 3–24.

[9] 高橋和子, 須山敦, 村山紀文, 高村大也, 奥村学. 2005. 職業コーディング支援システム（NANACO）の開発と JGSS-2003 における適用. 日本版 General Social Surveys 研究論文集［4］　JGSS で見た日本人の意識と行動, pp. 225–242.

[10] 高橋和子. 2008. 機械学習による ISCO 自動コーディング. 2005 年 SSM 調査シリーズ１２　社会調査における測定と分析をめぐる諸問題, pp.47–68.

[11] 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.

[12] 高橋和子, 魏大比, 田辺俊介, 吉田崇. 2012. 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 Web 公開. 言語処理学会第 18 回年次大会論文集, pp. 219–222.

[13] 田辺俊介. 2006. ISCO と SSM 職業分類の相違点の検討－国際比較調査における職業データに関する研究ノート－. 社会学論考　Vol.27, pp. 53–78.