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

高橋 和子
敬愛大学国際学部
takak@u-keiai.ac.jp

多喜 弘文
東京大学社会科学研究所
taki@iss.u-tokyo.ac.jp

田辺 俊介
早稲田大学文学学術院
tanabe.sh@waseda.jp

李 偉
東工大大学院理工学研究科
li.w.aa@m.titech.ac.jp

1　はじめに

社会調査において自由回答で得られた「職業」や「産業」の情報は, 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」）が必要である [3]. しかし, この作業は, 分類しなければならないコード数が多い上に, コード化のルールも複雑なため, サンプル数が多い場合は多大な労力や時間を要する. そこで, コーダの作業負担を軽減するために, 社会調査において国内標準となっている SSM 職業・産業コード [1] や, 近年, 必要性が高まってきた国際標準職業分類である ISCO（International Standard Classification of Occupations） [2] [1]を対象とした自動化システムを開発してきた [7, 9]. これらのシステムは, いずれも自然言語処理を基本にルールベース手法や機械学習（サポートベクターマシン；SVM）の適用により予測したコードを提示するもので, 特に初心者のコーダに対する支援システムとして有効であった.

自動化システムは, これまで主に JGSS（日本版 General Social Surveys）や SSM（Social Stratification and socila Mobility）調査などの大規模調査がおこなわれる際に利用されてきたが [8], それ以外にも調査者からの個別の依頼にも応じていたため, 依頼者, 開発者ともに時間や労力の面で問題があった. そこで, 社会調査をおこなう一般の研究者が広く利用できるものにするために, 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） [2]を窓口とする, Web を通じた利用のためのシステムへの改築に着手した [11].

[11] では, 運用担当者がシステムに関する専門的な知識をもたない場合でも容易に操作できるようにし, これまでの自動化機能を整理統合するとともに, 新たに, 自動コーディング結果に対する信頼性を示す 3 段階の確信度を付与する機能を追加した.

今回, さらに, 昨今の職業・産業コーディングを取り巻く状況に対応するため, 国際標準産業分類である ISIC（International Standard Industrial Classification of All Economic Activities）を付与する機能や, 過去の調査等ですでに付与された SSM コードがある場合に, この情報を用いて ISCO や ISIC を付与する機能の追加を行った. この自動化システムは, 昨年 11 月より SSJDA で試行提供が開始され, 利用者は SSJDA に申請書を提出し, 承認後に所定の形式のデータファイルをアップロードすれば, 最大 4 種類（SSM 職業・産業コード, ISCO, ISIC）のコーディング結果のファイルをダウンロードできる仕組みとなっている. 本稿では, 本システムの概要と公開方法, 運用について報告する.

以下, 次節では, 海外の職業・産業コーディングの自動化システムや公開方法について, 本研究との違いに注目して述べる. 3 節で本システムの概要, 4 節で本システムの公開と運用方法についてそれぞれ説明する. 最後にまとめと今後の課題について述べる.

2　関連研究

韓国では, 大韓民国統計庁の Web-based AIOCS（A Web-based Automated System for Industry and Occupation Coding）(Jung, Y. et al. 2008）があり, Web サイト上に, 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. 職業・産業コードは, いずれもレベル 4 までの階層構造をもつ ISCO や ISIC に基づくもので. それぞれ 442 個, 450 個である. 自動化の手法は, 本システムとは異なり, 処理時間の問題から, SVM ではなく, 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と組み合わせた計 6 種類を適用する. この中で正解率が最も高いのはルールベース手法, MEM, IRT を順に実行する場合で 76% である. これに対して本システムは, セキュリティ面からサーバに対する直接的なアクセスは行わず, また現時点ではデータの受け渡しはファイルのみである.

[1]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.

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

米国では, CDC（Centers for Disease Control and Prevention）の Web サイト上に SOIC（Standardized Occupation & Industry Coding）システムが公開され [3], 利用者はソフトウェアをダウンロードして処理を行う. 自動化の手法は単純で, 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, 両者とも正解は 63% である. これに対して本システムは, ソフトウェア構成が複雑なことや, 利用者のソフトウェア環境がさまざまに異なることからサポートが困難である点を考慮し, システム自体の公開は行わない.

CDC では, 2013 年に, SOIC の後継として新たに NIOCCS（The NIOSH Industry & Occupation Computerized Coding System）を公開し [4], 2000 年以前のセンサス・コードには SOIC. 2000 年以降のコードには NIOCCS を対応させている. NIOCCS はシステムを公開しておらず, 一問一答方式の他に, 入力情報は異なるものの, 本システムと同様にファイルによるデータの受け渡しを行う. また, 自動コーディングの結果に精度に関する確信度を 3 段階（High, Medium, Low）で付与する点も, 本システムと類似している.

3 職業·産業コーディング自動化システム

3.1 対象とするコードと自動化の手法

本システムが処理するコードを表 1, コードごとの自動化の手法を表 2 に示す.

自動化の手法において, SVM は, one-versus-rest 法により多値分類器に拡張した. 表中, 基本素性とは,「従業先事業の種類（自由回答）, 仕事の内容（自由回答）, 地位・役職（選択回答）」である. これ以外に ISCO と ISIC で用いる素性について述べる. まず, ISCO ではコードの決定に, 職業の遂行に必要なスキルレベル（＝教育·職業資格）が用いられるが, これは国際標準教育分類（ISCED）と対応し, 学歴を判断基準とするため [12], 学歴を追加した. また, ISCO においては, 第 1 位に予測された SSM 職業コードの追加が有効であったため [9], これを追加し, ISIC についても同様の対応を行った.

今回, 新たに ISIC を追加したことで, 現時点で多くの大規模社会調査で用いられている主要な職業や産業のコードに対応できるようになった. また, すべての処理に機械学習を適用したため, いずれのコードでも確信度を付与することが可能となった.

利用者は, 表 1 に示すコードのうち希望するコードを自由に選択できる. すべてのコードを処理する場合の手順を STEP 1～STEP 6 に示す.

[3] http://www.cdc.gov/niosh/soic/SOIC.About.html
[4] http://wwwn.cdc.gov/niosh-nioccs/

表 1: 対象とするコードの種類と個数

コードの種類	コード数	備考
SSM 職業（小分類）	約 200	
SSM 産業（大分類）	約 20	
ISCO（小分類）	約 400	階層構造（4 層）
ISIC（亜大分類）	約 60	階層構造（4 層）

表 2: 自動化の手法

コード	自動化の手法と用いる素性
SSM 職業	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	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	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	SVM（基本素性, SVM により第 1 位に予測された SSM 産業コード*）

＊：過去の調査等ですでに付与されたコードがある場合は, 予測コードではなく付与されたコードを用いる

STEP 1　職業・産業情報に対する形態素解析 [5]
STEP 2　ルールベース手法である ROCCO システム [6] の適用により, 仮 SSM 職業コードと仮 SSM 産業コードを出力
STEP 3　基本素性に, STEP 2 により出力された仮 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　基本素性に, STEP 2 により決定された仮 SSM 産業コードを追加して SVM を適用し, SSM 産業コードを決定
STEP 6　基本素性に, STEP 5 により決定された SSM 産業コード（第 1 位のみ）を追加して SVM を適用し, ISIC を決定

3.2 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 [5]. 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, 実験の結果, 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

3.3 システムの評価

本稿では, 事例に対して最終的に人手で付与されたコードを「正解」として扱う. ここでは, コードごとの正解率と確信度付与の有効性, 処理時間について報告する.

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, SSM コードでは, 訓練事例として JGSS の 2000 年〜2005 年調査データ（39,120 サンプル）, 評価事例として JGSS の 2006 年調査データ（JGSS-06）（2,203 サンプル）と 2005 年 SSM 調査データ（2005SSM）（16,089 サンプル）を用い, ISCO と ISIC では, 訓練事例として 2005SSM, 評価事例として JGSS-06 を用いた [6].

正解率

コード別の正解率を表 3 に示す. 表中, ISCO*と ISIC*は, 過去の調査等ですでに付与された正解 SSM 職業コード, 正解 SSM 産業コードをそれぞれ素性として用いた場合である（以下, 同様である）.

SSM 産業コードは, これまでルールベース手法のみを適用していたが, 今回, SSM 職業コードと同様に SVM との組み合わせを行った結果, 正解率が約 20% 向上した.

ISCO の正解率が低いため, 向上させる目的で, コード体系が階層構造であることを利用して, まず大分類（10 個）を学習した後に, 大分類ごとに小分類を学習する実験も行った. 第 1 位に予測されたコードについて, 大分類ごとの階層構造を利用した効果を調べた結果, 効果ありは 5 個, 効果なしは 3 個, 変化なしが 1 個であった [7]. また, 大分類ごとに正解率の高い方を選択して全体の正解率を算出したが, 本稿における手法（直接, 小分類を学習する）より 0.5% しか向上しなかった. これは, 大分類ごとに学習する場合に訓練事例のサイズがより小さくなり, カテゴリー数を減らした学習の効果を打ち消したためであると考えられる. 今後, ISCO コーディングが普及するにつれ, 訓練事例のサイズを拡大することが可能なため, 正解率の向上が見込める.

表 3: 正解率（第 3 位に予測されたコードまで含む）

コード	JGSS-06	2005SSM
SSM 職業	0.788	0.806
SSM 産業	0.908	0.916
ISCO	0.705	-
ISIC	0.801	-
ISCO*	0.748	-
ISIC*	0.862	-

表 4: 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)
SSM 産業	0.975(0.32)	0867(0.54)	0.437(0.14)
ISCO	0.963(0.05)	0.701(0.67)	0.276(0.28)
IISC	0.941(0.01)	0.919(0.56)	0.574(0.43)
ISCO*	0.947(0.05)	0.759(0.65)	0.300(0.30)
IISC*	1.000(0.01)	0.971(0.55)	0.671(0.44)

確信度の有効性

確信度別の正解率とカバー率（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）を表 4 に示す. NIOCCS では, High, Medium, Low をそれぞれ 90%, 70%, 30% としているが, 本稿では研究者を対象にするため, 確信度 A の正解率は高いほどよい. 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, ISCO や ISIC では非常に低かった.

処理時間

処理時間は PC の性能により異なるが, 約 2200 サンプル（JGSS-06）に対してすべてのコードを付与する場合, STEP 1 から STEP 6 にそれぞれ 0 分, 7 分, 34 分, 7 分, 13 分, 2 分（計 63 分）を要した.

4 本システムの公開と運用方法

本システムの利用者は, 所定の形式の入力用データファイル（学歴, 従業上の地位 · 役職, 仕事の内容, 従業

[5] 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は [10] により提案されているが, 本システムに組み込むことが困難であったため, 今回はこの手法を特徴づける「複数のスコア利用」を踏襲したより簡便な方法を提案した.

[6] 2005SSM と JGSS-06 以外の調査では, ISIC が付与されていない.

[7] 大分類が「0（Armed forces）」の場合は小分類が存在しない.

図 1: SSJDA 運用担当者操作画面

先の事業内容, 従業先の規模）[8]を準備し, （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.

（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（CSV 形式）を指定場所にアップロード
（4）［利用者］結果ファイル（CSV 形式）を指定場所からダウンロード

本システムが稼働するのは（3）と（4）の間で, 図 1 は, 本システムを稼働させたときに表示される初期画面である. 運用担当者は, 入力用データファイルを指定し, 必要なコードのチェックボックスをクリックすればよい. なお, セキュリティの点から, 運用担当者は利用者からのデータを e-mail 等では受けとらず, セキュアなオンラインストレージ構築パッケージ（Proself）を介する.

5 おわりに

本稿では. 現在 SSJDA で試行提供をおこなっている職業 · 産業コーディング自動化システムについて. その概要および公開と運用方法について述べた.

今後の課題は. システムの精度向上である. このための有効な対策の一つとして. 新たに正解が付与されたデータセットを既存の訓練事例に追加することがあるが. 将来的にはこの処理も自動化する必要がある.

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. 本研究は科研費（25380640）の助成を受けたものである.

[8]既存の SSM コードが存在し, ISCO や ISIC を希望する場合は, この SSM コードも入力する.

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

[3] 原純輔. 1984. 社会調査演習. 東京大学出版会.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 高橋和子, 田辺俊介, 吉田崇, 魏大比, 李偉 2013. Web 版職業 · 産業コーディング自動化システムの開発. 言語処理学会第 19 回年次大会論文集, pp. 769–772.

[12] 田辺俊介. 2008. SSM 職業分類と ISCO-88 の比較分析. 2005 年 SSM 調査シリーズ１　2005 年 SSM 日本調査の基礎分析－構造 · 趨勢 · 方法－, pp.31–45.