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)　多喜 弘文(東大)　田辺 俊介(早稲田大学)　李 偉(東工大)

研究の背景と目的

- 社会調査において**自由記述**により**職業・産業**情報を得た場合はコーディング作業 が必要
- コーダの負担を軽減するため**コーディング自動化システム**を開発
- 研究者が広く利用できるように、**Webによる一般公開**を行えるシステムとして再構築中

東大社会科学研究所SSJDAのWebサイトより試験的提供開始
(2103年11月より)

課題
(1) システムの精度向上
(2) 昨今の職業・産業コーディングを取り巻く状況に対応
(3) システム運用者の操作性向上

職業・産業コードの種類

	職業コード	産業コード
国内標準	SSM職業小分類(約200個) 例 501 自然科学研究者	SSM産業大分類(約20個) 例 10 農業　20 林業
国際標準	ISCO小分類(約400個) 例 1141 Senior official of political party organizations	ISIC亜大分類(約60個) 例 011 Growing of crops ; market gardening; horticulture

新規

SSMコード:1995年版
ISCO(International Standard Classification of Occupations):1988年版
ISIC (International Standard Industrial Classification of All Economic Activities):1988年版

入力データファイル例(CSV形式)

新規　過去の調査結果も利用可(この場合はISCOを出力)

NO.	学歴	地位＆役職	事業規模	勤務先事業の内容	仕事の内容	SSM職業コード
1	9	9	8	工場	コピー機のトナーカートリッジの製造	630
2	9	3	6	工場	ガラス吹き	625
3	11	4	9	福祉事務所	生活保護業務の現業員	654
4	11	8	8	予備校	事務	554
5	10	2	4	病院	看護士	514

選択肢(JGSS-2003配偶者に準拠)　自由記述

結果ファイル例(CSV形式)
(SSM職業コードの場合)

NO.	確信度	rank1	rank2	rank3
1	C	630	631	644
2	B	625	626	689
3	B	554	538	629
4	A	554	560	558
5	A	514	516	688

確信度　第1候補　第2候補　第3候補

確信度

自動コーディングの結果(がどの程度信頼できるかを機械学習により出力されたスコアに基づいて予測(3段階)

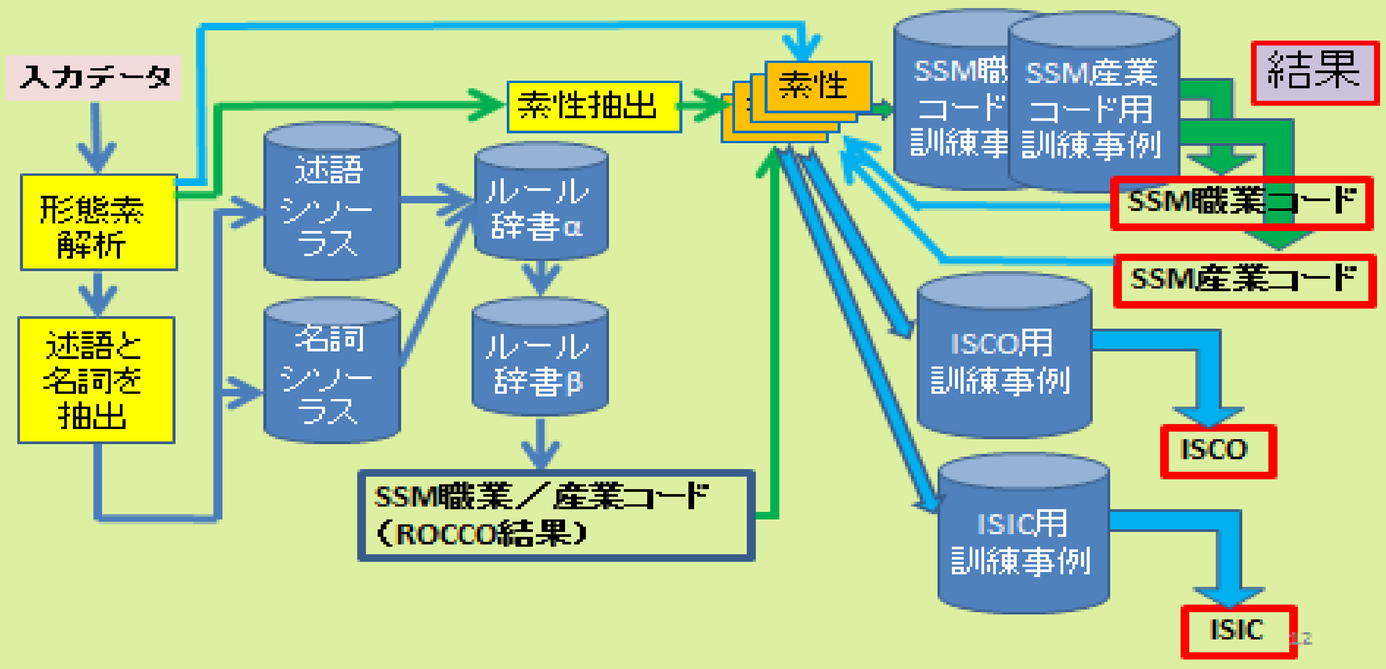
A : コーダの作業省略可能
B : できればコーダの作業必要
C : コーダの作業必要

複数のスコア利用

A: 第1位のスコア>0　第2位のスコア<=0
　第1位のスコアー第2位のスコア>α
B: 第1位のスコア>0　第2位のスコア<=0
　第1位のスコアー第2位のスコア<= α
C: A、B以外　　αは閾値

自動化システムの処理の流れ

ルールベース手法(ROCCO) → サポートベクターマシン



実験

[目的] システムの正解率と確信度の有効性をコードごとに調査
「正解」…事例に対して最終的に人手で付与されたコード

[評価尺度]
正解率 =正解した事例数／全事例数
カバー率=コードが付与された事例数／全事例数

[訓練事例]
国内標準コード用
JGSS-2000, -2001, -2002, -2003, -2005データセット(39,120サンプル)
国際標準コード用
2005SSMデータセット(16,089サンプル)

[評価事例]
JGSS-2006データセット(2,203サンプル)
2005SSMデータセット(国内標準コードの場合のみ)

(1)正解率(第3位まで)

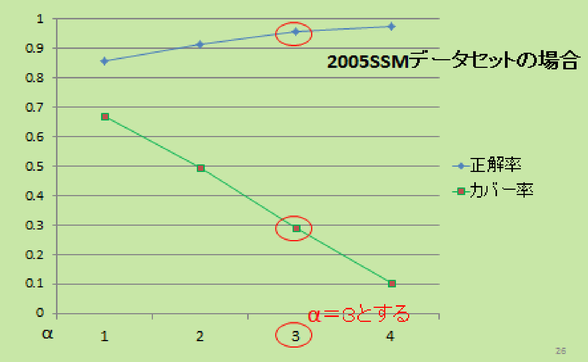
新規機能追加による結果　SVMの適用により約20%アップ

コード	JGSS-2006	2005SSM
SSM職業	78.8%	80.6%
SSM産業	90.8%	91.6%
ISCO	70.5%	−
ISIC	80.1%	−
ISCO*(正解SSM職業コード利用)	74.8%	−
ISIC*(正解SSM産業コード利用)	86.2%	−

職業コードの正解率が低い

(2)確信度の有効性

確信度Aの閾値αの変化による正解率とカバー率



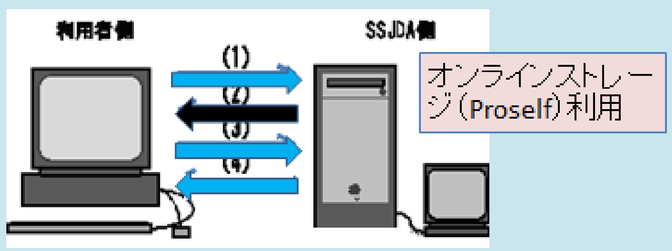
(2)確信度別の正解率とカバー率(α=3)

新規機能追加による結果　職業コードでも確信度Aの正解率は高い

コード	A	B	C
SSM職業	95.4%(29%)	71.6%(48%)	35.5%(23%)
SSM産業	97.5%(32%)	86.7%(54%)	43.7%(14%)
ISCO	96.3%(5%)	70.1%(67%)	27.6%(28%)
ISIC	94.1%(1%)	91.9%(56%)	57.4%(43%)
ISCO*	94.7%(5%)	75.9%(65%)	30.0%(30%)
ISIC*	100.0%(1%)	97.1%(55%)	67.1%(44%)

国際標準コードで確信度Aのカバー率が低い

利用手順

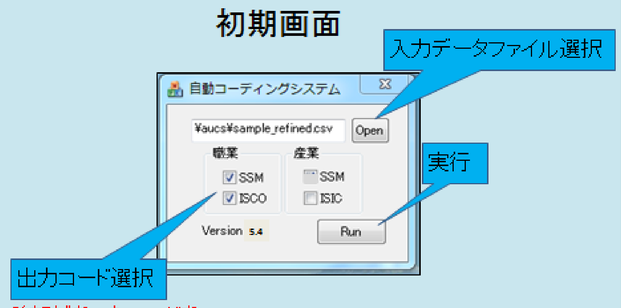


(1)[利用者] 利用申請書をメイルによりSSJDAに送信
(希望する職業・産業コードの種類を明記)
(2)[SSJDA] 利用申請書のチェック/承認
オンラインストレージ用IDとパスワードの発行
(3)[利用者] 入力データファイルをアップロード
(4)[利用者] 結果ファイルをダウンロード

[SSJDA]自動化システム

[SSJDA側]運用者操作画面

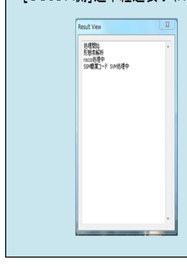
初期画面



希望があったコードを自由にチェック

[SSJDA側]途中経過表示
(SSM職業コード SVM処理中)

[SSJDA側]途中経過表示(Result View)



まとめ

(1) これまで開発してきた自動化システムはすべてSSJDAのWebサイトより一般公開
(2) さらに新機能として、「ISIC自動コーディング」「過去の調査等で付与済みの国内標準コードを利用した国際標準自動コーディング」を追加
(3) SSM産業コードは精度が向上したが、職業コードの精度は高くない
(4) 運用者にとって操作が容易なシステム

(1) 職業コードの精度向上
(2) 職業・産業コードの最新版(ISCO2008年版、SSM2015年版等)への対応
(3) システムの永続性の点から、メンテナンス処理の自動化

連絡先:高橋　和子(敬愛大学国際学部)　takak@u-keiai.ac.jp
本研究は平成25年度～平成27年度科研費(基盤研究(C)25380640)による成果の一部である