

職業・産業コーディング自動化システム

An Automatic Occupation and Industry Coding System

平成 25～27 年度 科学研究費補助金 基盤研究（C）

「社会調査の基盤を提供する自動コーディングシステムの Web 提供 :
その国際化と汎用化」

（課題番号 25380640）研究成果

研究代表者 高 橋 和 子

（敬愛大学国際学部 教授）

平成 28（2016）年 3 月

はしがき

本報告書は、現在、Web 版として公開されている職業・産業コーディング自動化システムについてまとめたものである。

社会学においては、職業や産業データは性別や年齢などと同様に重要な属性であり、正確を期する必要がある。このため、国勢調査でも行われているように、自由回答で収集したものを研究者自身の手で職業・産業分類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」とよばれるが、分類すべきコード（クラス）の数が非常に多いことやコード化のルールも複雑なことから、特に大規模調査の場合は多大な労力や時間を要するという問題を抱えている。また、多人数で長期間にわたる作業となるため、コーディング結果における一貫性の問題も指摘されている。

そこで、これらの問題を軽減する目的で、職業・産業コーディングを自動化するシステムの開発を行ってきた。開発は長期にわたり、その間、さまざまな自動化システムを構築してきたが、本報告書では、これらを統合した現行のシステムについて説明する。

本システムを利用してみたいという方は 1 節、システムの内容に関心がある方は 2 節と 3 節が参考になることと思う。本文で詳細まで説明できなかった部分を捕捉するため、資料編にいくつかの発表スライドを付けた。また、これまでの主な研究成果をテーマ別に掲載したので、関心のある方はこちらも参照していただきたい。

本システムは、ソフトウェア環境を整備してシステム本体をインストールすれば、利用者自身で実行することも不可能ではないので、参考までに 4 節で説明する。

システムの更新は今後とも必要になるが、これを容易にするため、作業の一部を自動化した。これについては 5 節で説明する。

本システムは 6 節に挙げるような課題があるものの、所期の目的は一応達成できたと考えている。次なる発展として、現在、一般の自由回答への拡張システムの開発に取り組んでおり、7 節でその概要について簡単に述べたい。

職業・産業コーディング自動化システムを多くの方に活用していただき、ご批判やご助言をいただければ幸いである。

2016 年 3 月

高橋 和子

目 次

はしがき

1	職業・産業コーディング自動化システムの概要と利用方法	1
1.1	システムの機能と性能	3
1.1.1	システムの機能	3
1.1.2	システムの性能	4
1.2	入力ファイル (CSV 形式)	4
1.3	結果ファイル (CSV 形式)	8
1.4	Web 公開版システム (試行提供中) の利用方法	9
2	システムの構成	10
2.1	システム構成図と基本的な処理の流れ	10
2.2	システムのファイル構造	10
2.2.1	lib フォルダ	11
2.2.2	data フォルダ	16
3	自動化のアルゴリズム	20
3.1	前処理	21
3.2	形態素解析 (juman)	22
3.3	ルールベース手法 (ROCCO)	22
3.3.1	三つ組みの抽出、ルールのマッチング、語の拡張	23
3.3.2	コードの修正	24
3.4	機械学習 (SVM)	24
3.4.1	素性の抽出	24
3.4.2	素性番号への変換	25
3.4.3	訓練事例による学習	25
3.4.4	未知の事例を分類	25
3.4.5	確信度の付与	25
3.5	後処理	26
4	システムの操作	27
4.1	システムの動作に必要なソフトウェア環境	27
4.2	インストールの方法	28
4.3	システムの実行方法	29
4.4	エラーで停止する場合と対応	31
5	システムの更新	32

5.1	訓練事例の更新方法（機械学習）	32
5.1.1	訓練事例追加の自動処理	32
5.1.2	訓練事例全体の差し替えまたは新規追加	33
5.2	シソーラスの更新方法（ルールベース手法）	34
5.3	ルール辞書の更新方法（ルールベース手法）	35
6	システムの課題と対応	36
7	自由回答一般への拡張可能性	39
	謝辞	40
	参考文献（資料編掲載以外）	40

資料編

(1)	『2005 年 SSM 日本調査コード・ブック』（2005 年 SSM 調査研究会編。2007 年）より 抜粋	
	・ 調査票の例（職業・産業情報部分）（p6）	資 1
	・ SSM 産業コード（大分類）（p84）	資 2
	・ SSM 職業コード（小分類）（p85～p87）	資 3
	・ 産業・職業のコーディング・ガイド（p88～p93）	資 6
	・ ISIC（p95～p96）	資 12
	・ ISCO-88（p97～p105）	資 14
	・ ISIC,ISCO のコーディング・ガイド（p106～p112）	資 23
(2)	The 9th Pacific Asia Conference on Knowledge Discovery and Data Mining （PAKDD）2005 発表スライド	資 30
(3)	The 11th Pacific Asia Conference on Knowledge Discovery and Data Mining （PAKDD）2007 発表スライド	資 36
(4)	独立行政法人統計センター招待講演スライド（2007 年 12 月 12 日）	資 43
(5)	The 6th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management（IC3K）2014 発表ポスター	資 55
(6)	言語処理学会第 22 回年次大会ワークショップ「言語処理の応用」発表スライド	資 56
(7)	システム開発に関連するこれまでの研究課題・研究組織と研究概要	資 64
(8)	システム開発に関連するこれまでの主要な研究成果（テーマ別）	資 66

あとがき

研究課題・研究組織

研究成果一覧

1. 職業・産業コーディング自動化システムの概要と利用方法

職業・産業コーディング自動化システムは、開発以来、精度向上や機能追加のために、アルゴリズムをさまざまに変えてきた。現行のシステムに至るまでの経緯について、最初に簡単に述べておく。

自動化システムは、当初、格フレームの概念を用いたルールベース手法の適用により、社会学において標準コードといえるSSM 職業小分類コードと産業大分類コードを付けるコードの支援システムとしてスタートを切った。このシステムはワークステーション上で開発したが、手軽に利用できるようにパソコン上に移植した。

その後、職業や産業情報のコーディングは、文章は短い、文書分類タスクとして扱うことができるのではないかと考え、機械学習の中でも分類性能の高さで評価されているサポートベクターマシン(SVM)を適用し、さらに、SVMとルールベース手法を組み合わせた手法を開発した。

一方で、社会学における国際比較研究の隆盛に対応するため、処理の対象とするコードを、ILOにより定められた国際標準職業分類ISCO (International Standard Classification of Occupations) や、国際標準産業分類ISIC (International Standard Industrial Classification of All Economic Activities) とするシステムも開発した。対象としたコードは、社会学で用いられるISCO (小分類) とISIC (亜大分類) である。

SSM 職業・産業コードは、もともと1968年版ISCO やISIC を源とする日本標準職業・産業分類を社会学で使用しやすいように改変されたものであったが、その後のISCO とISICにおける改訂の結果、両者の対応関係は複雑化した。これが、ISCO やISIC のために新規に自動化システムを開発した理由である。なお、時間の関係上、ISCO やISIC については、ルールベース手法は構築しなかった。

自動化システムでは、いずれも入力ファイル (CSV 形式) にある職業や産業情報から予測されたコードをCSV形式で提示するため、コードはこれを参考にしながらコーディングを行うことができる。特に初心者のコードに対する有効性が評価され、二次分析のための大規模調査であるJGSS (Japanese General Social Surveys ; 日本版総合的社会調査) において、初回の2000年以降、利用されてきた。また、10年ごとに実施されるSSM (Social Stratification and social Mobility) 調査 (社会階層と社会移動全国調査) においても、2005年調査に引き続き、2015年調査でも利用された。SSM 調査は、社会学の中でも職業や産業データがとりわけ重要な役割を果たす階層移動研究の調査で、大規模である上に、本人の初職から現職にいたるまでの職業や産業の履歴に加え、配偶者、父親、母親についての職業も収集されるため、コーディング作業量の問題が特に大きい。

自動化システムを利用することにより、コードの作業内容は楽になったが、すべてのデータに対してコーディング作業を行うことについては変わらない。そこで、コードの作業の絶対量についても軽減できるように、自動コーディング後にコードの作業が必要かど

うかを示す目安として、3 段階（A:不要、B:できれば要。C:必要）の確信度を付与する機能を追加した。

このように、自動化システムは機能を充実させながら、主として大規模調査で利用されてきたが、これ以外にも、研究者個人やグループからの依頼を受けて開発者自身が処理を行ってきたケースがある。そこで、次には、システム実行者（開発者）・利用者（研究者）双方が使いやすいシステムについて検討を始めた。

その結果、システム実行者のためには、これまでコードの種類ごとに、いくつかのプログラムが独立に開発されてきたものを整理・統合し、容易に処理が行えるようなシステムとして再構築することにした。また、システム利用者のためには、Web を通じて自由に利用できる仕組みを検討し、利用者自身が入力データのファイルをアップロードすれば、結果をダウンロードできる方法を考案した。現在、自動化システムは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（CSRDA）に置かれ、CSRDA によりWeb を通じた利用サービスが試行提供されている。

以上のように、自動化システムは変遷を重ねながら、コードの支援という当初の目的を一定程度達成できたと評価できよう。現行のシステムについて、まず、本節でWeb公開版の概要と利用方法を説明し、2節、3節、4節で詳細について述べる（図1-1参照）。

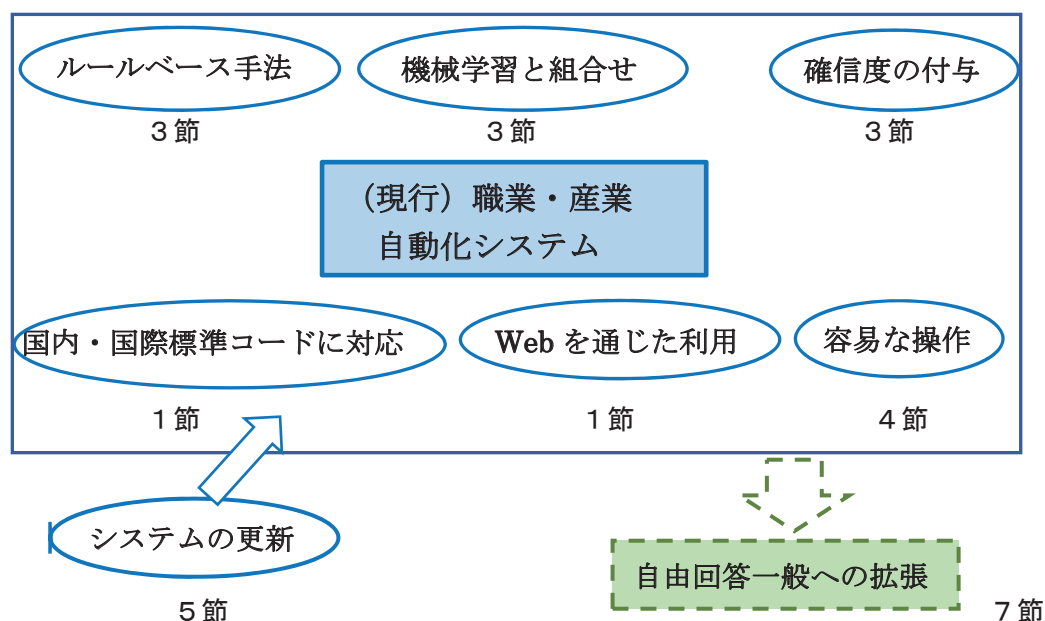


図 1-1 現行の職業・産業コーディング自動化システムの概要とその拡張

自動化システムの次なる段階は、永続的な利用のために、開発者以外でもシステムの更新を行うことができることである。公開版には組み込んでいないが、このための処理を一部開発している（5節を参照のこと）。さらに、自動化システムの汎用的利用に向けては、自由回答一般に拡張したシステムの開発にも取り組んでいる（7節を参照のこと）。

1.1. システムの機能と性能

1.1.1. システムの機能

システムが変換するコード体系は、次の4種類である。

- | | |
|------------------|------------------------|
| ・ SSM 職業コード（小分類） | 国内の社会調査で標準的に用いられる職業コード |
| ・ SSM 産業コード（大分類） | 国内の社会調査で標準的に用いられる産業コード |
| ・ ISCO（小分類） | ILOにより定められた国際標準の職業コード |
| ・ ISIC（亜大分類） | ILOにより定められた国際標準の産業コード |

図 1-2 システムが変換するコード体系

このうち、SSM 職業（小分類）と SSM 産業（大分類）には、表 1-1 に示すように、当初はなかった 700 番台、800 番台の新規コードが追加されているため、現行のシステムでは扱うようにしている。各コードの詳細については、資料編(1)を参照していただきたい。

表 1-1 対象とするコードの種類とコードの数

	コードの種類	コードの数	備考
国内 標準	SSM 職業コード (小分類)	約 200	1995 年版が基本であるが、700 番台、800 番台もあり
	SSM 産業コード (大分類)	約 20	1995 年版が基本（卸売、小売、飲食店を区別）
国際 標準	ISCO（小分類）	約 400	ISCO-88 階層構造 4 桁すべて利用
	ISIC（亜大分類）	約 60	ISIC-Rev3 階層構造上位 2 桁まで利用

利用者はこの中から自由にコードを選ぶことができる。例えば、SSM 職業コード（小分類）だけを選んでよいし、これと SSM 産業コード（大分類）の2種類でもよい。また、すべてのコード（4種類）を選んでよい。

ISCO や ISIC に変換する場合、過去の調査などですでに SSM 職業／産業コードが付いている場合は、これを正解として利用すると精度が高まることがわかっている（表 1-2 参照）。本システムではこのようなケースも想定し、入力ファイルに、SSM 職業／産業コードの正解も追加することで対応している（1.2 節の（2）⑥⑦を参照のこと）。本報告書では、このような場合を ISCO*や ISIC*と表記することにする。

1.1.2. システムの性能

本報告書では、正解率を「システムが正解した事例数／全事例数」とする。ただし、職業・産業コーディングにおける「正解」は、「調査実施者が最終的に与えたコード」を指すため、正解率は、正解とされたコードとの「一致率」であるともいえる。

システムが候補として提示する第3位までのコードについての正解率を表1-2に示す。コードの種類によって多少の違いはあるが、おおざっぱに言えば、国内標準コードでは、職業が約80%、産業が約90%であり、国際標準コードでは、これより約10%ずつ低下している。しかし、国際標準コードでは、すでに付与されている国内標準コードを与える、と、約5%ずつ高くなる。

表 1-2 コードの種類別正解率

コードの種類	第3位まで
SSM 職業コード	約 80%
SSM 産業コード	約 90%
ISCO	約 70%
ISIC	約 80%
ISCO*	約 75%
ISIC*	約 86%

本システムで一度に処理できるのは5,000事例である。入力ファイルがこれより大きな場合は、数回に分けて処理する必要がある。5,000事例を処理する時間は、コンピュータの性能や訓練事例のサイズによっても異なるが、パソコンでは約2～3時間程度である。

1.2. 入力ファイル（CSV形式）

システムで扱うデータは、入・出力ともファイルに限定しており、一問一答形式のものは扱わない。これは、調査終了後のアフターコーディングを想定するためである。

入力ファイルの内容は、「職業や産業に関する情報」と「学歴」である。形式については、CSRDAのWebに掲載されている「入力ファイルの形式」（<http://csrda.iss.u-tokyo.ac.jp/autocode-form.pdf>）に詳しい。ISICの処理も行うことができること等、若干の補足と本報告書に合わせた形式への改変を行った上で、ここに掲載させていただく。

（1）入力ファイルの全体の概要

自動コーディングシステムは、以下の形式に則った、1行が一つの職業（あるいは産業）分のデータとなるように入力されたCSV形式のファイルが必要である（図1-3、図1-

4 参照)。ただし、そのファイルでは「項目名」は付けず、1 行目からデータを入力する。G 列や H 列は、以下の説明にあるように特別な場合以外には使用しない。

A 列	: 通し番号 (数字)	半角
B 列	: 学歴 (選択回答)	半角
C 列	: 従業上の地位・役職 (選択回答)	半角
D 列	: 産業 (従業先事業の種類) ** (自由回答)	全角
E 列	: 職業 (仕事の内容) * (自由回答)	全角
F 列	: 従業先の規模 (選択回答)	半角
G 列	: SSM 職業コード*** (数字)	半角
H 列	: SSM 産業コード**** (数字)	半角

図 1-3 入力ファイルの形式

- ・通し番号は必須。
- ・*印の項目は職業コードには必須。
- ・**印の項目は産業コードには必須。
- ・*印のない項目は必須ではないが、自動コードの回答精度に影響する。
- ・自由回答に、「。」 (全角ピリオド)、「.」 (半角ピリオド)、「全角空白」、「?」、「〒」、「☆」などの特殊文字は、処理上問題になるため、削除する。
- また、回答の最後に「。」を付けないこと (途中に入っていても問題ない)。

悪い例：保険会社の事務。

この他に自由回答でエラーとなる例を 4.4 節で説明しているので、参照していただきたい。

- ・***印の項目は、利用者側ですでに SSM 職業コードを与えたデータに対し、新たに ISCO-88 コードを付与したい場合に必須。
- ・****印の項目は、利用者側ですでに SSM 産業コードを与えたデータに対し、新たに ISIC-Rev3 コードを付与したい場合に必須。

A 列	B 列	C 列	D 列	E 列	F 列	G 列	H 列
11	10	9	保険会社の支店	保険会社の事務	11		
12	12	8	会社の売店	販売員	8		
62	9	7	夫の社会保険事務所	社会保険事務所の総務、経理	3		
289	9	10	農業	野菜を作っている	2		
465	9	1	訪問介護事業	訪問介護の経営、介護福祉士	4		
1093	13	14	無回答	営業 (外回り)	13		

図 1-4 入力ファイルの例

（２）各項目の説明

①A 列： 通し番号

一つの職業に一つの通し番号が必要となる。回答者 1 人に対して複数の職業についての回答がある場合、それぞれの職業について別のファイルとして作成する（例： 回答者 1 人に現職と初職の回答がある場合、現職用ファイルと初職用ファイルの 2 つを用意する）。

②B 列： 学歴の選択肢

学歴についてはシステム上、日本版総合社会調査（JGSS）で用いられた以下のコードを用いて処理している。そのため、他の選択肢で行った調査データの場合は、以下のコードに合わせてリコードする。

1	旧制尋常小学校（国民学校を含む）	8	新制中学校
2	旧制高等小学校	9	新制高校
3	旧制中学校・高等女学校	10	新制短大・高専
4	旧制実業学校	11	新制大学
5	旧制師範学校	12	新制大学院
6	旧制高校・旧制専門学校・高等師範学校	13	わからない
7	旧制大学・旧制大学院		

③C 列： 従業上の地位・役職の選択肢

従業上の地位や役職についてはシステム上、以下のコードを用いて処理している。他の選択肢を用いて行った調査の場合は、以下のコードに合わせてリコードする。

1	経営者・役員	8	臨時雇用・パート・アルバイト
2	常時雇用の一般従業者 役職なし	9	派遣社員
3	常時雇用の職長、班長、組長	10	自営業主・自由業者
4	常時雇用の係長、係長相当職	11	家族従業者
5	常時雇用の課長、課長相当職	12	内職
6	常時雇用の部長、部長相当職		
7	常時雇用であるが、役職はわからない	14	わからない

④D 列： 産業（従業先事業の種類）と E 列： 職業（仕事の内容）

回答の自由記述の内容を「全角（のみ）」で入力する。またシステム処理の関係上、以下の諸点に注意する。

- ・半角文字を含めない。
- ・英字は大文字にする。

- ・空白（含「全角空白」）やピリオド（.）、特殊な記号（?、〒、→、@、☆ など）を含めない。

以上の問題があるファイルは適切に処理できないため、特に注意すること。ただし、回答の長さは処理に直接は影響しないため、特に配慮する必要はない。

⑤F 列： 従業先の規模（企業規模）の選択肢

従業先の規模（企業規模）についてはシステム上、日本版総合社会調査（JGSS）で用いられた以下のコードを用いて処理している。そのため、他の選択肢で行った調査データの場合は、以下のコードに合わせてリコードする（※）。

1	1 人	8	500～999 人
2	2～4 人	9	1,000～1,999 人
3	5～9 人	10	2,000～9,999 人
4	10～29 人	11	1 万人以上
5	30～99 人	12	官公庁
6	100～299 人	13	わからない
7	300～499 人		

※）実施された調査の選択肢が、JGSS よりも粗い企業規模カテゴリとなっているデータの場合（たとえば「10～99 人」という選択肢がある場合など）、上記のコードをそのまま当てはまることができない。その場合の対処としては、次のような方法が考えられるが、どのような方法が適切かは、研究目的や内容によるかと思われるので、利用者の方で適宜判断する。

- ・自動コーディングシステムが職業コードを付与する際に、「30 人以上の規模の企業か、以下の企業か」で管理職の処理が異なり、「1 人、5 人未満、30 人未満、100 人未満、官公庁」という区分が、仮コード修正に一定程度用いられるので、この情報を考慮の上、どのようにコードを与えるべきかを判断する。
- ・機械的に中央値で置き換える（10 人～99 人の場合は、 $(10+99) \div 2 \div$ 「55 人」→「30～99」に含める）等。

⑥G 列： SSM 職業コード

利用者側で SSM 職業コードをすでに付与しているデータに対して、ISCO コードを新たにコーディングしたい場合に記入する。不明や無回答の場合は「999」と記入し、無回答のないようにする。

⑦H 列：SSM 産業コード

G 列と同様に、利用者側で SSM 産業コードをすでに付与しているデータに対して、ISIC コードを新たにコーディングしたい場合に記入する。不明や無回答の場合は「999」と記入し、無回答のないようにする。

1.3. 結果ファイル（CSV 形式）

結果ファイルの例を図 1-5 に示す。出力内容は、「通し番号」の付いたサンプルごとに、「第 1 位に予測された結果（rank1）、第 2 位に予測された結果（rank2）、第 3 位に予測された結果（rank3）」の計 3 個の候補となるコードで、第 1 位の候補に対しては、「確信度」も付与する。候補の順位は、機械学習が出力するスコアの大きい順である。

A 列	B 列	C 列	D 列	E 列
ID	確信度	rank1	rank2	rank3
1	C	630	631	644
2	B	624	626	689
3	B	554	538	629
4	A	554	560	558
5	A	514	516	688

図 1-5 結果ファイルの例（SSM 職業コードの場合）

確信度は、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したものである。本システムでは次の 3 段階としている。

- A : コーダの作業は不要
- B : コーダの作業はできれば要
- C : コーダの作業が必要

コーダの作業の絶対量を削減したいときにもっとも有用な指標は、確信度 A が付与された場合の正解率である。これは、4 種類のコードのいずれも、どのような実験においてもつねに 94%以上を示しており、一応、信頼できる値となっている。ただし、確信度 A が付与される事例の割合は、国内標準コードでは約 30%であるが、国際標準コードは 5%未満であり、国際標準コードにおける改善が必要である。確信度 B や確信度 C の正解率はバラツキがあるが、おおむね確信度 B の場合は 70%程度、確信度 C の場合は約 45%程度である。

1.4. Web 公開版システム（試行提供中）の利用方法

本システムは、現在、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（CSRDA）の Web（<http://csrda.iss.u-tokyo.ac.jp/joint/autocode/>）を通じて利用できる。

CSRDA に申請した書類（「自動コーディング（職業・産業）システム利用申請書」）が受理されれば、だれでも利用できる。申請書類に記載する主な内容は、「調査名」と「希望するコードの種類」であるが、提供されたデータは学術目的にのみ利用することを誓約事項としている。

本システムの利用者は、1.2 節で説明した形式の入力データファイルを準備し、CSRDA から指定された場所にアップロードすれば、本システムが実行され、希望する職業や産業のコーディング結果をダウンロードできる仕組みとなっている（図 1-6 参照）。

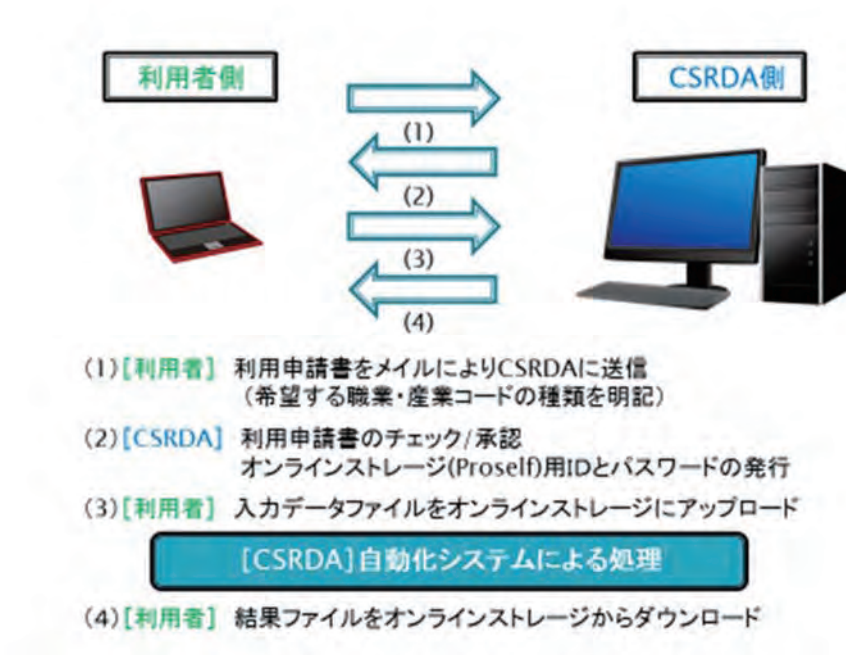


図 1-6 Web 公開版システム（試行提供中）の利用方法

本節では、システムを外側から眺めた場合の説明を行った。次節以降では、システムの内部についてより詳細に説明する。

2. システムの構成

2.1. システム構成図と基本的な処理の流れ

本システムの構成図を図 2-1 に示す。矢印は処理の流れで、基本的には、「前処理 → 形態素解析 → ルールベース手法 → 機械学習 → 後処理」である（処理内容の詳細については 3 節で説明する）。

図中、左半分はルールベース手法に関わる部分で、右半分は機械学習（SVM）に関わる部分である。両者を組み合わせることにより、システムの精度向上が実現された。

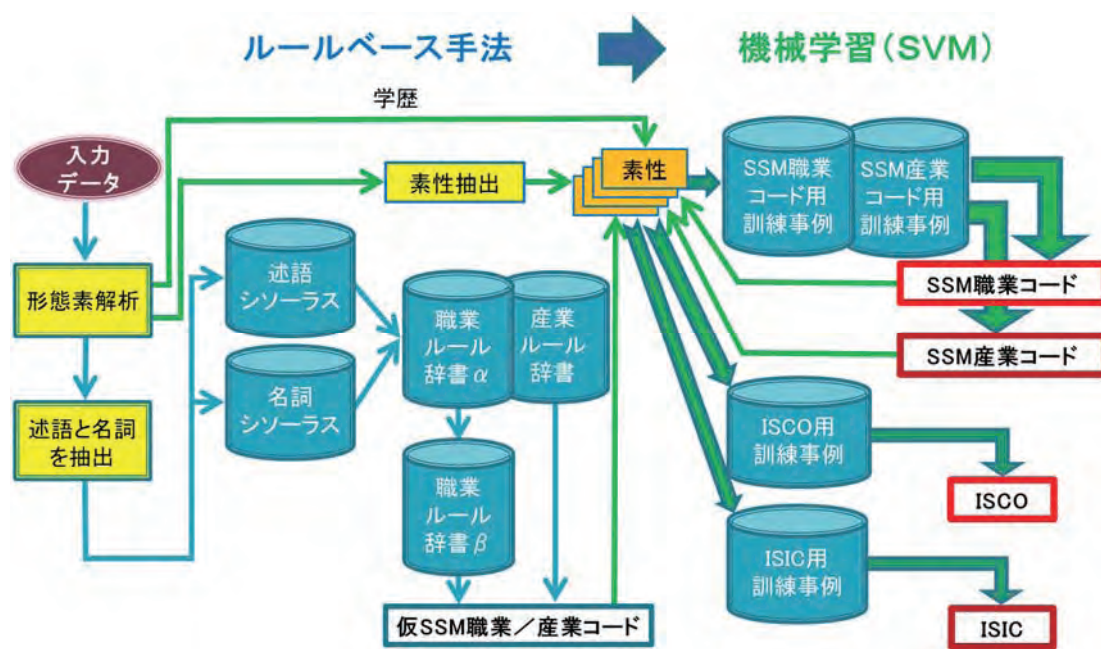


図 2-1 システム構成図と処理の流れ

2.2. システムのファイル構造

システムの本体は、C ドライブ直下にある aucs フォルダである。aucs フォルダには、図 2-2 に示すように、システムの実行プログラム (aucs*.exe) と 2 つのフォルダ (lib フォルダ、data フォルダ) が存在する。

実行プログラム aucs*.exe の*はバージョン情報で、実際のプログラム名は「aucsV4.4.exe」や「aucsV7.4.exe」などである。aucs*.exe は本システムのメインプログラムで、lib フォルダにある各種プログラム等をコントロールする。ソースコードは C++ 言語であるが、バイナリパッケージとして置かれている。

lib フォルダには、システムで用いる各種プログラムや SVM で必要となる訓練事例などがある。

data フォルダには、ルールベース手法の結果のファイルやシステムの最終的な結果ファイルを保存する report フォルダ、SVM 処理途中の結果を保存する result フォルダ（処理終了後にファイルを削除）、SVM 処理のために必要な temp フォルダの 3 つのフォルダと、ルールベース手法で用いられるシソーラス（2 種類）とルール辞書（2 種類）がある。

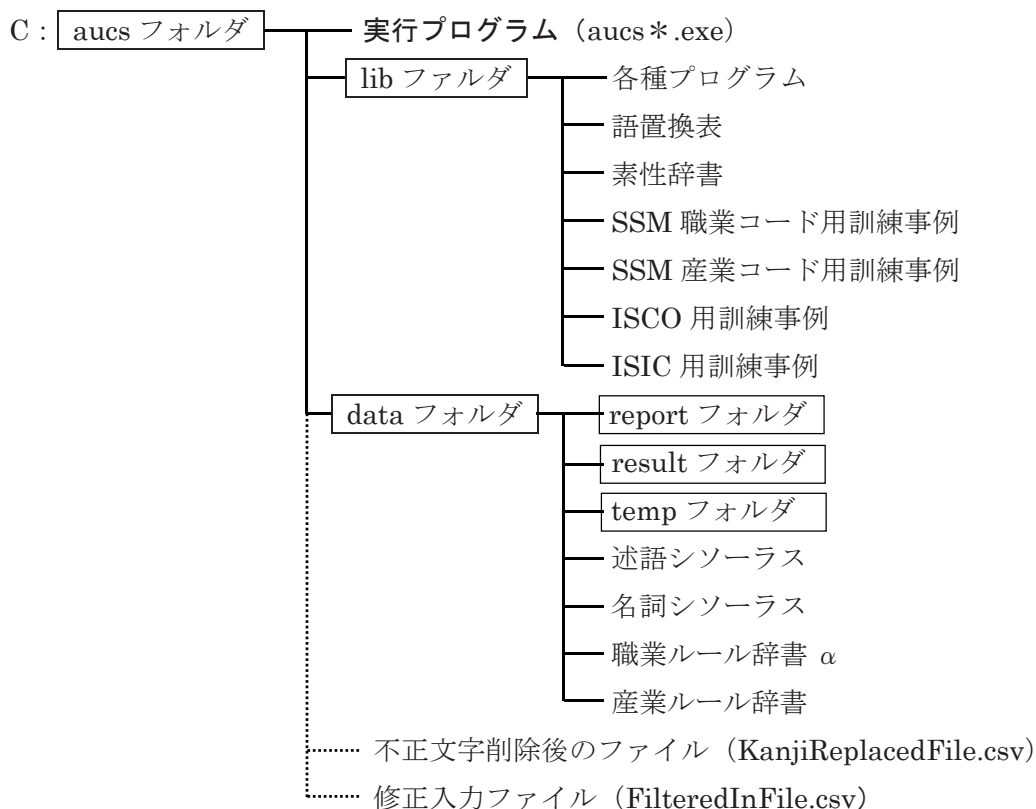


図 2-2 aucs フォルダのファイル構造（処理開始時点）

システムを稼働させるためには、aucs フォルダ以外に、各種プログラムのためのコンパイラ等のソフトウェアがすべて C ドライブ直下に置かれていなければならない。これについては、4.1 節で説明する。

入力ファイルは C ドライブにある必要はなく、どの場所にあってもよい。システム操作画面で指定できる（図 4-3 参照）。入力ファイルは、前処理により不正文字の削除等の修正や語の置換が行われ、C ドライブ直下に修正入力ファイルとして新規に生成される。このファイルが自動化処理の対象となる。

2.2.1. lib フォルダ

（1）各種プログラム

プログラムには、表 2-1 に示すように、コンパイルが必要なファイルと、そのまま実行

できる exe ファイルが混在する。コンパイルが必要なものは、C ドライブ直下にコンパイラがインストールされている必要がある（インストールの方法は 4.2 節で説明する）。

表 2-1 lib フォルダ中の各種プログラム

処理内容	開発言語	備考
ルールベース手法（ROCCO）	LISP	exe ファイル（実行モジュール）
SVM 実行のための素性を生成	Perl	コンパイラが必要
SVM 実行	Ruby	コンパイラが必要
操作画面による入力処理、前処理等	Java	コンパイラが必要

（２）語置換表（replace_ttable_Input.csv）

語置換表は、職業や産業の分類という点では同一視してもよいと考えられる語の対応表ファイル（CSV 形式）で、A 列の語が B 列の語に置換される。例えば、出現語における表記の揺れの解消（表 2-2、表 2-3 参照）、近年増えてきたカナカナ語への対応（表 2-4 参照）や、素性辞書（次項目で説明）にすでに登録されている語と同一視できる語への対応（表 2-5 参照）である。

語置換表の目的は、異なり語を無用に増やさないことで SVM の処理における素性空間を大きくしないためだけでなく、形態素解析で形態素を適切に切り出せないという失敗を減らすための対策である。

現時点での語置換表の見出し語（行数）は、1,481 語である。

表 2-2 表記の揺れの例（その 1）

A 列	B 列
センバン	旋盤
せん盤	旋盤
セン盤	旋盤
旋ばん	旋盤

表 2-3 表記の揺れの例（その 2）

A 列	B 列
ケアマネージャ	ケアマネージャー
ケアマネージャー	ケアマネージャー
ケアマネージャー	ケアマネージャー
ビルメンテ	ビルメンテナンス

表 2-4 カタカナ語の対応例

A 列	B 列
グリーンアテンダント	乗務員
メディカルセンター	病院

表 2-5 登録語への変換例

A 列	B 列
助言	アドバイス
シンキン	信用金庫

語置換表が適用されると、例えば、「ビルメンテ」は「ビルメンテナンス」となるため、もともとあった「ビルメンテナンス」は「ビルメンテナンスナンス」となってしまうという問題がある。このため、語の置換が終了した時点で、例えば、「メンテナンスナ

ス」は「メンテナンス」と置換する必要がある。このため、語置換表の最後に表 2-6 に示すような語の対応を置き、語が置換された場合でも最終行まで検索を行っている。

表 2-6 最終解決のための対応例

A 列	B 列
メンテナンスナンス	メンテナンス
員員	員

(3) 素性辞書 (svmdicuniqrd)

本システムで使用している SVM では、素性としては「語」ではなく、「番号」が用られるため、語と品詞の組（以下、単語と呼ぶ）を番号（素性番号）に変換する必要がある。素性辞書はこのための対応表である（表 2-7 参照）。

現時点での素性辞書の見出し語は 15,089 語である。素性辞書にない語はすべて「200,000」なる番号に変換される。このため、適宜、更新を行わないと、新しく出現した語は、違う語であっても同じ番号に変換されるため、精度の低下を招く。この対策として、Version.8.1（未公開版）では、自動的に更新を行う機能をもつ（5 節を参照のこと）。

表 2-7 単語と素性番号の対応例

よい	形容詞	3	S E	未定義語	361
を	助詞	33	アフターサービス	未定義語	384
及び	助詞	34	製造	名詞	2432
作る	動詞	205	電気機器	名詞	2818

(4) 訓練事例

SVM における学習のために、表 1-1 に示す 4 種類のコード別に訓練事例が存在する。本システムでは、訓練事例を、信頼性の高い JGSS データセットと 2005SSM 調査データセットの最終結果（正解）のみを用いて生成している。

訓練事例の構成を図 2-3 に示す。素性はすべて番号で、辞書順にソートされている。

正解 素性 1 : 素性 1 の頻度 素性 2 : 素性 2 の頻度 . . . 素性 n : 素性 n の頻度
--

図 2-3 訓練事例の形式（1 事例分）

本システムでは、「仕事の内容」に出現した語は、素性辞書で変換された素性番号をそのまま素性として用いるが、「従業先事業の種類」に出現した語は、素性番号に「20,000」をプラスした番号を素性とすることで、両者の回答を区別している。

表 2-8 に示すように、訓練事例に用いる素性はコードの種類により多少異なるため、以

下ではそれぞれについて説明する。

① SSM 職業コード用訓練事例

SSM 職業コード用訓練事例の素性は、「基本素性」と「ルールベース手法の結果（SSM 職業コード）」で構成される。本報告書では、基本素性とは、「従業上の地位・役職」（選択回答）「仕事の内容」（自由回答）「従業先事業の種類」（自由回答）の3つを指す。

選択肢の番号を素性辞書にある番号と区別するために、「従業上の地位・役職」は、回答された選択肢番号に「500,000」をプラスした番号を素性とする。

また、同様の理由で、「ルールベース手法の結果（SSM 職業コード）」は、出力された SSM 職業コードに「1,000,000」をプラスした番号を素性としている。

図 2-4 に示す訓練事例の例は、素性番号から、「従業上の地位・役職」が「8」（臨時雇用・パート・アルバイト）で、「仕事の内容」に、「1234」（一般）、「2016」（事務）が各 1 回、「従業先事業の種類」に、「91」（社）、「275」（（：左かっこ）、「276」（）：右かっこ）、「3349」（旅行）、「3350」（旅行業）が各 1 回出現しており、「ルールベース手法の結果」が「554」（総務・企画事務員）であることがわかる。この事例には、正解として「554」が付けられている。

554 1000554:2 1234:1 200091:1 200275:1 200276:1 2016:1 203349:1 203350:1 500008:1
--

図 2-4 SSM 職業コード用訓練事例の例

現時点での SSM 職業コード用訓練事例は、正解付きの JGSS データセットと 2005SSM 調査データセットから生成された 49,794 事例である。

② SSM 産業コード用訓練事例

SSM 産業コード用訓練事例の素性は、「基本素性」と「ルールベース手法の結果（SSM 産業コード）」で構成される。

ルールベース手法の結果（SSM 産業コード）は、出力された SSM 産業コードに「2,000,000」をプラスした番号を素性とする。

図 2-5 に示す例は図 2-4 と同じ事例に SSM 産業コードの正解が付与されたものである。素性から、この事例の「ルールベース手法の結果」が「82」（旅行業）であることがわかる。この事例には、正解として「82」が付けられている。

82 1234:1 2000082:1 200091:1 200275:1 200276:1 2016:1 203349:1 203350:1 500008:1

図 2-5 SSM 産業コード用訓練事例の例

現時点での SSM 産業コード用訓練事例は、SSM 職業コード用訓練事例と同一の事例から生成された 49,794 事例である。

③ ISCO 用訓練事例

ISCO 用訓練事例の素性は、「基本素性」と「SVM により第 1 位に予測された SSM 職業コード」、「学歴」で構成される。学歴を用いる理由は、ISCO-88 の決定にはスキルレベルが考慮されが、これをデータとして特には収集されないため、学歴を代用できる変数としたためである。

SVM により第 1 位に予測された SSM 職業コードは、第 1 位に予測された SSM 職業コードに「8,000,000」をプラスした番号を素性としている。

学歴は、選択肢に「600,000」をプラスした番号を素性とする。

図 2-6 に示す訓練事例の例は、素性から、学歴は「9」（新制高校）であり、SVM により第 1 位に予測された SSM 職業コードが「550」（会社・団体等の管理職員）であることがわかる（「仕事の内容」や「従業先事業の種類」については説明を略する）。この事例には、正解として、「4121」（ACCOUNTING AND BOOK-KEEPING CLERKS）が付けられている。

4121 1370:1 1476:1 1649:1 19:1 200025:1 200034:1 201339:1 202432:1 203038:1 25:1 2726:1 500006:1 600009:1 8000550:1
--

図 2-6 ISCO 用訓練事例の例

現時点での ISCO 用訓練事例は、正解付きの 2005SSM データセットから生成された 16,088 事例である。

④ ISIC 用訓練事例

ISIC 用訓練事例の素性は、「基本素性」と「SVM により第 1 位に予測された SSM 産業コード」で構成される。

SVM により第 1 位に予測された SSM 産業コードは、第 1 位に予測された SSM 産業コードに「9,000,000」をプラスした番号を素性としている。

図 2-7 に示す訓練事例の例は図 2-6 と同じ事例に ISIC の正解が付与されたものである。素性から、この事例の「SVM により第 1 位に予測された SSM 産業コード」が「60」（製造業）であることがわかる（ISIC の素性には学歴を用いない）。この事例には、正解として、「36」（Manufacture of furniture）が付いている。

36 1370:1 1476:1 1649:1 19:1 200025:1 200034:1 201339:1 202432:1 203038:1 25:1
2726:1 500006:1 9000060:2

図 2-7 ISIC 用訓練事例の例

現時点での ISIC 用訓練事例は ISCO 用訓練事例と同一の事例から生成された 16,088 事例である。

SVM の処理において、どの素性がどのような値となって、どのコードに用いられるのかを表 2-8 にまとめておく。

表 2-8 素性の番号生成方法と訓練事例で用いられるもの（○印）

用いる素性	用いる番号	SSM 職業	SSM 産業	ISCO ISCO*	ISIC ISIC*
仕事の内容に出現した語	素性辞書の素性番号	○	○	○	○
従業先事業の種類に出現した語	素性辞書の素性番号に 200,000 をプラスした番号	○	○	○	○
従業上の地位・役職	選択肢に 500,000 をプラスした番号	○	○	○	○
ルールベース手法の結果 (SSM 職業コード)	出力コードに 1,000,000 を プラスした番号	○			
ルールベース手法の結果 (SSM 産業コード)	出力コードに 2,000,000 を プラスした番号		○		
学歴	選択肢に 600,000 をプラスした番号			○	
<u>SVM により第 1 位に予測された SSM 職業コード</u>	出力コードに 8,000,000 を プラスした番号			○	
<u>SVM により第 1 位に予測された SSM 産業コード</u>	出力コードに 9,000,000 を プラスした番号				○

2.2.2. data フォルダ

(1) report フォルダ

report フォルダはもっとも重要なフォルダで、希望したコードについての結果ファイルが CSV 形式で保存される。また、現行のシステムでは中間結果となっているが、ルールベース手法による最終結果も CSV 形式で保存される。これは、ルールベース手法の結果をチェックするのに役立つ。

① 結果ファイル

1 行目に項目名があり、2 行目以降に、SVM による最終結果と、第 1 位の予測結果に対する確信度が表示される（図 1-5 参照）。

② ルールベース手法の最終結果ファイル

ルールベース手法では、自由回答から、(述語、表層格、名詞) の三つ組みを抽出し、その各々に対して、次項で説明するルール辞書を検索し、マッチしたルールのコードを付ける (3.3 節を参照のこと)。このため、出力されるコードの数は、抽出された三つ組みの数により異なる。例えば、図 2-8 において、ID が 1101 の事例は 4 個、1103 の事例は 3 個のコードが付けられている。「999」(不明) は、マッチするルールがない場合に付けられるコードである。

```
1101  704  704  704  672
1103  688  999  592
```

図 2-8 ルールベース手法の最終結果の例

(2) result フォルダ

SVM による途中結果 (クラスごとに出力された事例のスコア) を保存するが、処理終了後に削除される。

(3) temp フォルダ

SVM の処理のために必要なフォルダである。

(4) シソーラス

三つ組みとして抽出されるすべての述語と名詞に対してルールを生成するのは不可能である。そこで、それぞれを抽象化 (一般化) したレベルの語でルールを生成しておき、シソーラスとして、実際に自由回答に出現するレベルの語や、形態素解析により切り出される語 (形態素) に展開した語との対応付けを行うためのファイルである。

① 述語シソーラス (jyutsugo.txt)

語や品詞が異なっても、職業や産業の分類という点では同一視した方がよいと考えられる述語や述語相当語 (サ変名詞など) には、同じ述語コードを付ける。例えば、「作る」「製造」「製作」は同一の語として、同一の述語コード「386 1」を付ける。

述語シソーラスは、1 行が(よみ 原形 述語コード)で構成されるテキスト形式のファイルである。図 2-9 に例を示す。原形は、juman により切り出された形態素の原形である。juman では、カタカナ語は、複合語であっても、一続きのカタカナ部分すべてが一語とし

て切り出される場合が多い。

現時点での述語シソーラスの見出し語は 10,871 語で、述語コードは 2,880 個ある。

よみ	原形	述語コード
↓	↓	↓
(すいみんぐいんすとりくたあ	スイミングインストラクター	241 22)
...		
(せいさく	制作	386 1)
...		
(せいぞう	製造	386 1)

図 2-9 述語シソーラスの一部

② 名詞シソーラス (meishi.txt)

名詞シソーラスは、1 行が(ルール辞書で用いられる語 具体的なレベルの語 1 ... 同 n)で構成されるテキスト形式のファイルである。図 2-10 に例を示す。

ルール辞書には「自動車 1」はあるが、「自動車」以下の語はない。ルール辞書で用いられる語はどのような語であってもよいが、具体的なレベルの語は、実際に juman で切り出される語でなくてはならない。

ルール辞書で 用いられる語	具体的なレベルの語
↓	↓
(自動車 1	自動車 車 カー 乗用車 バス タクシー ハイヤー トラック・・・)

図 2-10 名詞シソーラスの一部

現時点での名詞シソーラスの見出し語は、330 語である。

(5) ルール辞書

ルール辞書は、SSM 職業コードや産業コードを決定するために、述語コードごとにルールを記述したテキスト形式のファイルである。その形式は図 2-11 に示すとおりで、見出しの述語コードから決定される職業／産業コードと表層格、名詞の情報が繰り返される。

((述語コード)(SSM 職業／産業コード 1 (表層格 名詞 1 1 ... 名詞 1 n ₁))
...
((SSM 職業／産業コード m (表層格 名詞 m 1 ... 名詞 m n _m)))

図 2-11 ルール辞書の形式

① 職業ルール辞書 α (syokugyo.txt)

職業ルール辞書 α の例を図 2-12 に示す。1 行目は、述語コード「386 1」に対して、表層格「を」、名詞「ソフトウェア」、「システム」、「ウェブページ」があれば、SSM 職業コードが「506」（情報処理技術者）となるルールである。述語コードが同じでも、表層格や名詞が違えば、SSM 職業コードが異なることがわかる。最終行は、表層格や名詞が存在しない場合（抽出に失敗した場合も含む）のルールで、SSM 職業コードが「704」（製品製造作業：作っている製品が明記されていない場合）となる。

述語コード	SSM 職業コード	表層格	名詞
↓	↓	↓	↓
((386 1)	(506	(を	ソフトウェア システム ウェブページ))
	(507	(を	放射線装置 1 耐火物))
	...		
	(704	(0))	

図 2-12 職業ルール辞書 α の一部

現時点での職業ルール辞書 α の見出し語（異なる述語コード）は、4,224 語である。すべての述語コードがルール辞書に出現するわけではない。

② 産業ルール辞書 (Sangyo.txt)

産業ルール辞書の例を図 2-13 に示す。職業ルール辞書 α の場合と同じ述語コード「386 1」の例を掲載した。この述語コードの場合、表層格と名詞の違いにより、SSM 産業コードが「30」（漁業）や「160」（法律・会計サービス業）などとなる。最終行のルールにあるように、表層格や名詞が存在しない場合（抽出に失敗した場合も含む）は、SSM 産業コードが「60」（製造業）となる。

述語コード	SSM 産業コード	表層格	名詞
↓	↓	↓	↓
((386 1)	(30	(で	船内))
	(160	(を	税務書類))
	...		
	(60	(0))	

図 2-13 産業ルール辞書の一部

現時点での産業ルール辞書の見出し語（異なる述語コード）は、948 語である。すべての述語コードがルール辞書に出現しないのは、産業ルール辞書でも同様である。

3. 自動化のアルゴリズム

自動化の処理の流れの基本は、「前処理 → 形態素解析 → ルールベース手法 → 機械学習 → 後処理」であるが、表 3-1 や図 3-1～図 3-3 に示すように、コードの種類や入力データの形式により多少異なる。

ISCO／ISIC には、SSM 職業／産業コードと異なり、このコードのためのルールベース手法は存在しないが、機械学習において、SVM により第 1 位に予測された SSM 職業／産業コードも素性とするため、このコードを得るために、SSM 職業／産業コードのためのルールベース手法は適用される（図 3-2 参照）。

ただし、過去の調査などですでに SSM 職業／産業コードの正解があり、これを利用者が入力ファイルで与える場合には、この正解を用いることができるために、ルールベース手法を適用する必要はない（図 3-3 参照）。

表 3-1 コードの種類別自動化の手法

コードの種類	自動化の手法（機械学習で用いる素性）
SSM 職業コード	ルールベース手法の結果を機械学習（SVM）に組み込む （基本素性、ルールベース手法の結果）
SSM 産業コード	ルールベース手法の結果を機械学習（SVM）に組み込む （基本素性、ルールベース手法の結果）
ISCO	機械学習（SVM） （基本素性、 <u>SVM により第 1 位に予測された SSM 職業コード</u> 、学歴）
ISIC	機械学習（SVM） （基本素性、 <u>SVM により第 1 位に予測された SSM 産業コード</u> ）
ISCO*	機械学習（SVM） （基本素性、利用者が与えた正解 SSM 職業コード、学歴）
ISIC*	機械学習（SVM） （基本素性、利用者が与えた正解 SSM 産業コード）

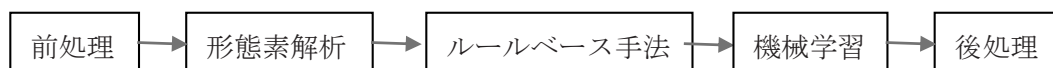


図 3-1 SSM 職業／産業コードの場合

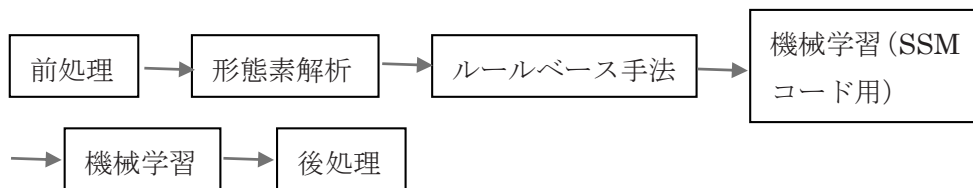


図 3-2 ISCO/ISIC の場合

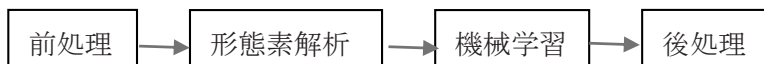


図 3-3 ISCO*/ISIC* の場合 (SSM 職業/産業コードの正解が入力された場合)

3.1. 前処理

前処理では、入力ファイル（図 1-3 参照）の自由回答中に含まれる「全角空白」や不正文字を削除した後、同一視できる語をまとめる処理を行う。

（１）空白や不正文字を削除

本システムでは、「全角空白」をデータの区切りとしているため、自由回答中にこれが含まれると、「従業先事業の種類」「仕事の内容」「従業先の規模」のデータの位置を正しく認識できないために、エラーで停止してしまう。

不正文字はエラーの原因となったり、「述語、表層格、名詞」の三つ組み抽出の失敗となる可能性が高いため削除しておく（結果は KanjiReplacedFile.csv に保存される）。不正文字の具体例は、1.2 節（１）を参照していただきたい。

（２）同一視できる語を置換

出現した語ごとに、語置換表の A 列を 1 行目から順に検索し、該当する語があれば B 列の語に置換する（表 2-2～表 2-6 参照）。置換後も最後の行まで検索する。

入力ファイルに対する前処理が終了すると、最終的に、修正版入力ファイル（FilteredInFile.csv）が aucs フォルダ直下に生成される。以下の処理は、この修正版入力ファイルに対して行われる。



図 3-4 ファイルでみる前処理の処理過程

3.2. 形態素解析 (juman)

本システムは、形態素解析を行うプログラムとして、京都大学長尾研究室で開発された juman3.6.1 を使用する。

juman は、コマンドラインからも実行できるが、本システムでは、入・出力をいずれもファイルとしている。入力ファイルは、3.1 節で生成した修正版入力ファイルである。出力ファイル (jgondo.txt) は、この後の処理であるルールベース手法の入力ファイルとなるため、日本語 EUC コードのファイル (jgondo) に変換される。これは、ルールベース手法をワークステーション上で開発したためである。

本システムでは、「-e オプション」(完全な形態素情報を文字とコードで表示) を使用する。例として、図 1-4 に示した入力ファイルの例のうち、通し番号 62 の E 列(「仕事の内容」)「社会保険事務所の総務、経理」の形態素解析結果を示す。

形態素	よみ	原形	品詞	詳細な品詞
↓	↓	↓	↓	↓
社会	しゃかい	社会	名詞 6	普通名詞 1*0*0
保険	ほけん	保険	名詞 6	普通名詞 1*0*0
事務所	じむしょ	事務所	名詞 6	普通名詞 1*0*0
の の の		助詞 9	接続助詞 3*0*0	
総務	そうむ	総務	名詞 6	普通名詞 1*0*0
、 、 、		特殊 1	読点 2*0*0	
経理	けいり	経理	名詞 6	サ変名詞 2*0*0
EOS				

図 3-5 juman (-e オプション付き) による形態素解析結果の例

形態素解析の結果のうち、ルールベース手法で用いるのは、「原形」と「詳細な品詞」である。形態素を適切に切り出せるか否かは、ルールベース手法に始まるこの後の処理の成否に大きく影響する。

修正版入力ファイル (FilteredInFile.csv) → jgondo.txt (→ jgondo)

図 3-6 ファイルでみる juman の処理過程

3.3. ルールベース手法 (ROCCO)

本システムのルールベース手法は、「述語、表層格、名詞」の三つ組構造により格フレ

ームの概念を利用する点に特徴がある。簡単にいえば、「職業／産業コードを決定するルールをこの三つ組みにより生成しておき、回答から抽出した三つ組みとマッチしたルールがあれば、そのコードに決定する」というアイディアであるが、これを考案した過程は次のとおりである。

まず、「職業や産業の情報は、大まかには動作として捉えることができる」との考えにより、自由回答の中から動作を表すものとして、品詞が動詞である語を抽出する。しかし、動作は動詞だけではなく、例えば「製造」のように、サ変名詞によっても表現される。一方で、職業や産業の情報は「医師」や「デパート」のように名詞として出現する場合もある。そこで、動作を広く解釈することにし、文末にあつて、品詞が動詞、サ変名詞、普通名詞である語を述語相当語（以後、述語とよぶ）として捉える。

職業や産業を分類する際の決め手となる情報は、述語によって異なっている。例えば、職業の場合、述語が「製造」であれば「どこで」ではなく「何を」、「教える」であれば「何を」ではなく「どこで」が必要な情報となる。このため、抽出された述語に対しては、必要な表層格と名詞を調べる必要があり、これが回答から抽出した「述語、表層格、名詞」の三つ組とマッチすれば、このルールのコードに決定できることになる。

LISP 言語により開発したこの自動化システムは、ROCCO (Rule based OCCupation C0ding) と名付けられ、長い間利用された。現行のシステムにも組み込まれている。

3.3.1. 三つ組みの抽出、ルールのマッチング、語の拡張

日本語では、述語は文末に来ることが多いため、抽出は文の最後から開始する。例えば、図 3-5 の例では、まず、最後に位置するサ変名詞の語「経理」が抽出される。

次に、抽出された述語を「述語シソーラス」により述語コードに変換する。例えば、「経理」は「371 3」と変換される。

さらに、この述語コードにより職業ルール辞書 α を検索し、回答から生成した三つ組みとマッチするルールがあるかをチェックする。例えば、図 3-7 に示すように、述語コードが「371 3」の場合は、職業ルール辞書 α により、「を 公認」が回答中から抽出できれば、SSM 職業コードは「519」（公認会計士、税理士）となり、何も抽出できなければ「559」（会計事務員）となる。図 3-5 の例では、回答に「を 公認」がなく、名詞シソーラスで「公認」の拡張もできないため、何も抽出できないことになり、「559」のコードに決定される。

((371 3) (519 (を 公認))
(559 ()))

図 3-7 職業ルール辞書 α における述語コード「371 3」の三つ組み

図 3-5 の例では、「経理」の前に並列表現を表す「、」があるため、さらに抽出が続け

られ、その前にある普通名詞である「総務」が抽出される。「総務」は述語コード「243 9」が付けられ、職業ルール辞書 α により SSM 職業コードが「554」に決定される。

図 3-5 では、「総務」の前に、「の」「事務所」「保険」「社会」が存在するが、述語コード「371 3」も「243 9」も、SSM 職業コードの決定に表層格「の」は不要であるため、これらの語は抽出されない。結局、図 3-5 の例では、読点「、」で切られた 2 つの三つ組みが並列表現として抽出され、それぞれに対して SSM 職業コード「559」「554」が決定されることになる。

SSM 産業コードにおいても、職業ルール辞書 α を産業ルール辞書に代えて、同様の処理が行われる。

3.3.2. コードの修正

ここまでは、自由回答の情報からのみコードを決定するアルゴリズムを説明した。しかし、SSM 職業コードでは、例えば、管理職や建設関係等の場合、「従業上の地位」「従業先の規模」「学歴」の情報により、コードを修正する必要もある。そこで、これをチェックするルールを職業ルール辞書 β (図 2-1 参照) として用意し、職業ルール辞書 α の後に利用した結果をルールベース手法としての最終コードとする (結果は rocco_occcode.txt に保存される)。

なお、職業ルール辞書 β はファイルではなく、LISP プログラム中に記載されている。

jgondo → rocco_occcode.txt → roccoresult.txt → occcode.csv
--

図 3-8 ファイルでみるルールベース手法 (ROCCO) の処理過程
(SSM 職業コードの場合)

3.4. 機械学習 (SVM)

本システムは、機械学習として、サポートベクターマシン (SVM) を適用した教師付き学習を適用する。学習に有効な素性選択を行う実験の結果、表 3-1 に示す素性を用いている。

3.4.1. 素性の抽出

まず、回答から素性を抽出するが、表 3-1 に示したように、コードの種類により用いる素性が多少異なる。本システムでは、選択肢も自由回答と同様に形態素解析を行い、各コードに応じた素性を抽出する。このとき、訓練事例と同じ種類の素性を抽出する。

本システムの機械学習では、ルールベース手法による結果も素性として用いる点に特徴がある。ルールベース手法と機械学習の組み合わせ方はさまざまに考えられるが (資料(4)

を参照のこと)、実験の結果、もっとも精度が高かったこの方法を採用した。

3.4.2. 素性番号への変換

抽出した素性は素性番号に変換する。変換の仕方は表 2-8 に示したとおりである。次に、素性番号ごとに出現回数を数え、図 2-3 に示す形式のものを生成するが、「正解」はすべて「999」としておく。最後に、素性番号を辞書順にソートする。

現行のシステムでは、素性辞書にない語の素性番号はすべて「200,000」となるが、Version.8.1（未公開版）では、入力ファイルに対する「前処理」段階で、新出語には自動的に素性番号を付けて素性辞書に追加する。また、訓練事例の追加のときも同様に、正解付き事例において新規の語があれば、自動的に素性辞書に追加する（5.1.1 節を参照のこと）。いずれの場合も、処理の途中で再度その語が出現したときは、素性辞書に追加された素性番号に変換される。この更新機能の追加により、システムの精度低下を軽減する効果が期待できる。

3.4.3. 訓練事例による学習

自動コーディングを行うコードの種類に応じて、SSM 職業コード訓練事例、SSM 産業コード訓練事例、ISCO 訓練事例、ISIC 訓練事例のいずれかを選んで学習を行う。

3.4.4. 未知の事例を分類

表 1-1 に示すように、職業・産業コーディングは数十または数百のコード（クラス）に分類する多値分類である。しかし、SVM は 2 値の分類器であるため、本システムでは、one versus rest 法により多値分類器に拡張する。

one versus rest 法では、未知の事例が、あるクラスに属するか否かという 2 値分類をすべてのクラスに対して行い（結果は result.txt に保存される）、最終的にスコアのもっとも大きいクラスに分類する方法である。例えば、SSM 職業コードの場合は約 200 個のクラスがあるため、事例ごとに約 200 個のクラスについて判定を行い、最終的なクラスを決定することになる。

```
svminhg&roccoresult.txt → svma_addocc → svma_addocc_sort → result.txt
```

図 3-9 ファイルでみる SVM における処理過程（SSM 職業コードの場合）

3.4.5. 確信度の付与

第 1 位に予測されたクラスに対する確信度の各段階を決める条件は、図 3-10 のとおりである。このとき、第 1 位に予測されたコードのスコアだけでなく、第 2 位に予測されたコードのスコアも用いることで、第 1 位に予測されたコードに対する確信度 A の正解率を向上させる効果が期待できる。これは、第 1 位に予測されたコードに対するクラス所属確

率を推定する際、複数の分類スコアを利用することで、推定の精度を高めることができたという実験結果にヒントを得ている（資料（3）を参照のこと）。

α は閾値で、実験の結果、本システムでは $\alpha=3$ としている。

A : 第 1 位のスコア > 0 第 2 位のスコア ≤ 0

第 1 位のスコア - 第 2 位のスコア $> \alpha$

B : 第 1 位のスコア > 0

第 1 位のスコア - 第 2 位のスコア $\leq \alpha$

C : A、B 以外の場合

図 3-10 確信度 A、B、C の決定条件

3.5. 後処理

後処理では、事例ごとに各クラスのスコアを調べ、第 1 位の候補のコードとしてはもっとも大きな値をもつクラス、第 2 位の候補のコードとしては 2 番目に大きいクラス、第 3 位の候補のコードとしては 3 番目に大きいクラスを選び、図 1-5 に示す CSV 形式の結果ファイルを生成する。その際、図 3-10 で決定される確信度を B 列に付与して完成させる。

結果ファイル名は、SSM 職業コードは SSM_syo_trust.CSV（図 3-11 参照）、SSM 産業コードは SSM_san_trust.CSV、ISCO は ISCO_trust.CSV、ISIC は ISIC_trust.CSV である。

result.txt → 結果ファイル (SSM_syo_trust.CSV)

図 3-11 ファイルでみる後処理の処理過程（SSM 職業コードの場合）

4. システムの操作

1.4 節で説明したように、現在、本システムは Web 公開版として CSRDA により試行提供中であり、利用申請が承認されれば、利用者は入力ファイルを指定の場所にアップロードするだけで結果が得られる仕組みとなっている。

しかし、システムの操作自体は容易であるため、もし利用者側でソフトウェア環境を整えることができ、システムのバージョンアップやエラー時の対応などのサポート体制を必要としないということであれば、利用者自身で稼働させることも不可能ではない。

そこで、参考までに、システムの稼働に必要なソフトウェア環境やそのインストールの方法、システムの操作方法について、本節で説明を行うことにする。

4.1. システムの動作に必要なソフトウェア環境

システムは、1 節で述べたように、開発当初から現在の構成で設計されていたわけではなく、アルゴリズムの改善や機能の追加が次々に行われてきたという経緯がある。また、操作を容易にするために、これまで別々に開発していたシステムを整理・統合したため、例えば、開発に用いられたプログラム言語も多数混在し、全体としてやや複雑なソフトウェア環境となっている。

本システムが動作するためには、次の (a) ～ (c) が利用できる環境である必要がある。開発言語としては、他にも C 言語や LISP を用いているが、これは実行モジュールにして置いている。

- (a) 日本語形態素解析用ソフト (juman)
- (b) Java、Perl、Ruby のプログラム言語
- (c) Windows 上での linux コマンド

インストールが必要なライブラリは、図 4-1 に示すようなリリースフォルダ (libsoft フォルダ) として用意している (現在は非公開)。libsoft フォルダには、システム本体 (aucs フォルダ) も含めている。

• aucs フォルダ	システム本体
• juman フォルダ	日本語形態素解析用ソフト juman 本体
• Ruby192 フォルダ	Ruby のコンパイラなど
• ActivePerl-5.14.2.1402-MSWin32-x86-295342.msi	Perl コンパイラのインストーラー・パッケージ
• cygwin フォルダ	Windows 上で通常の linux のコマンドを利用可能にする

図 4-1 リリースフォルダの内容

4.2. インストールの方法

libsoft フォルダの内容は、STEP1 によりインストールできるが、インストール後に、STEP2 で path の設定を更新しておく必要がある。以下の説明は、Windows 7 における例である。

STEP1 次の4つのフォルダをインストールまたはコピーする

(1) juman・・・libsoft/juman

- ① juman フォルダを C ドライブの直下にコピー
- ② juman フォルダの「juman.ini」を C:¥windows にコピー
- ③ juman フォルダの「cygwin1.dll」を C:¥windows にコピー
- ④ Windows スタートボタン→「アクセサリ」→「コマンドプロンプト」で DOS 画面を表示して juman フォルダに移動する（例 `cd ../../juman`）。
- ⑤ 「C:¥juman>」が表示されるので、「makedic.bat」と入力して実行（数秒かかる）
- ⑥ DOS 画面を閉じる
- ⑦ juman フォルダを C:¥Program Files フォルダにコピー

(2) Ruby・・・libsoft /Ruby192

Ruby192 フォルダを C ドライブの直下にコピー

(3) Perl・・・libsoft /ActivePerl-5.14.2.1402-MSWin32-x86-295342.msi

- ① 「ActivePerl-5.14.2.1402-MSWin32-x86-295342.msi」をダブルクリック
- ② 指示に従って了解する旨のボタンを押していくと、C ドライブの直下に Perl フォルダができる

(4) cygwin フォルダを C ドライブの直下にコピー・・・libsoft/cygwin

STEP2 path の設定を更新する

- ① Windows スタートボタン→「コンピューター」を右クリックし、「プロパティ」をクリック
- ② 画面左側にある「システム詳細設定」をクリック
- ③ 環境変数ボタンをクリック
- ④ 「システム環境変数」欄の「path」をダブルクリック
- ⑤-1 「変数値」の末尾に「;」を入力し、「C:¥aucs¥lib」を追加入力
- ⑤-2 同様に続けて「;C:¥Ruby192¥bin」を入力
- ⑤-3 同様に続けて「;C:¥Perl¥bin」を入力
- ⑤-4 同様に続けて「;C:¥cygwin¥bin」を入力
- ⑥ 「OK」ボタンを押して表示画面を終了していく

4.3. システムの実行方法

システムの実行方法は、以下に示すとおりである。

STEP1 aucs*.exe をダブルクリックし、初期画面を表示させる。*はバージョン情報である。



図 4-2 システム操作初期画面

STEP2 Open ボタンを押して入力ファイルを指定する。図 4-3 の例では、入力ファイルが aucs フォルダの直下にあるが、どこにあっても指定できる。



図 4-3 入力ファイルの指定

STEP3 変換を希望するコードにチェックをする。

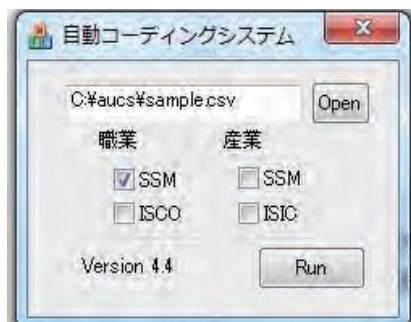
図 4-4 は、1 種類のコードを希望した場合の例である。複数ある場合は、該当するコードのすべてにチェックできる。4 種類すべてにチェックしてもよい。

SSM 職業コードが付与されたデータを ISCO に変換したい場合 (ISCO*) は、STEP2 で入力データファイルの G 列に付与済みの SSM 職業コードを入れたファイルを入力ファイルとして指定し、ISCO にチェックする。

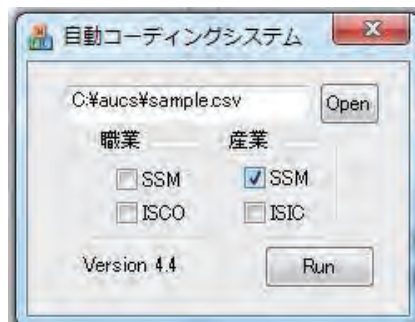
SSM 産業コードが付与されたデータを ISIC に変換したい場合 (ISIC*) は、STEP2 で入力データファイルの H 列に付与済みの SSM 産業コードを入れ

たファイルを入力ファイルとして指定し、ISIC にチェックする。

① SSM 職業コードに変換したい場合



② SSM 産業コードに変換したい場合



③ ISCO に変換したい場合



④ ISIC に変換したい場合

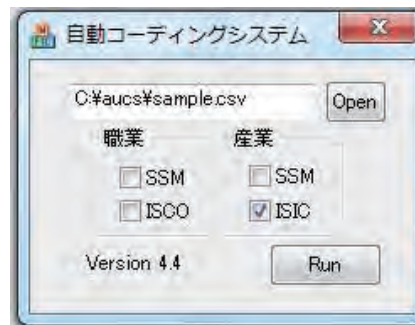


図 4-4 変換を希望するコードの種類にチェック

STEP4 Run ボタンを押すと、画面に処理の過程が表示されながら（図 4-5、図 4-6 参照）、処理が行われていく。



図 4-5 処理途中のメッセージ画面推移（SSM 職業コードと ISCO にチェックした場合）

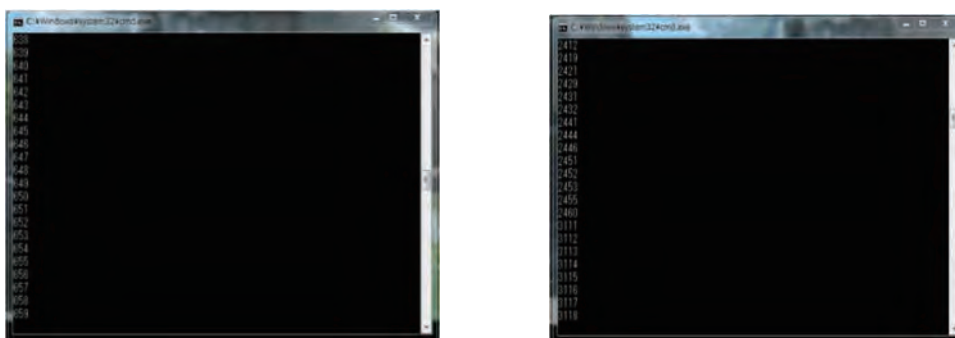


図 4-6 処理途中の画面例（SSM 職業コードと ISCO にチェックした場合）

4.4. エラーで停止する場合と対応

これまで、システムがエラーで停止することはほとんどなかったが、次の場合は確実にエラーとなるので、注意していただきたい。いずれもルールベース手法の過程で起きる。

- ① 1.2 節で述べたように、回答に「全角空白」が混入している場合である。現行のシステムでは、前処理段階でこのチェックを行って削除しているが、画面に、「データの型が合わない」旨のエラーメッセージが表示された場合は、念のために回答をチェックしていただきたい。
- ② ルールベース手法において、三つ組みの抽出は 6 組までしかできないため、並列表現が 7 組以上ある（とシステムが判断した）場合に起きる。例えば、「きゅうり、大根、キャベツ、かぼちゃ、人参、玉ねぎ、ナスを作っている」なる回答は、システムは、文末から、（作る を ナス）、・・・、（作る を 大根）と 6 個の三つ組みを抽出し、7 個目を抽出しようとした時点でエラーとなる。この場合、画面に、どの事例のどこの場所でエラーになったかが表示されるため、回答を確認し、結果に影響がなさそうな語（例えば、「ナス」）を削除することで対応する。
- ③ 具体的な例であるが、回答を形態素解析した結果、「書」という語が一語で切り出され、かつそれが「、」「。」の直前にある場合に起きる。例えば、juman では、「納品書」「企画書」「受領書」のように、「書」の前にある語が単独に存在し得る確率が高いと計算されれば、それぞれ「納品」「企画」「受領」と「書」のように 2 つの形態素として切り出される（「秘書」「司書」のように、「書」を含めて一語となる確率が高い場合は問題ない）。このような現象が起きる理由は判明しないが、文字コードの問題である可能性が考えられる。本質的な解決法ではないが、このような回答があった場合は、「書」を「書類」または「書き」などに置換することで対応する。

ここまで、現行のシステムとして Web 公開版システムについて説明した。次節では、現時点では未公開であるが、システムの更新機能について説明する。

5. システムの更新

これまでも職業・産業コーディングを取り巻く状況は変化してきた。本システムに関連するところだけでも、新しい内容の仕事が登場することで、自由回答に新規の語が登場する。職業や産業コードも、既存のものから分化させる場合が多いとはいえ、新規のものが作られる。さらには、コード体系そのものが定期または不定期に更新される等である。これは今後も続くため、システムの更新作業は必須である。本節では、開発者以外でもこれを行うことができるよう、訓練事例やシソーラス、ルール辞書の更新について説明する。

5.1. 訓練事例の更新方法（機械学習）

訓練事例の更新が必要になるのは、次の2つの場合である。一つは、調査後のコーディング作業が終了して最終的なコードが決定され、正解付き事例が蓄積されるが、これを現行の訓練事例に追加したい場合である。もう一つは、コード体系の変更により、訓練事例全体を変更せざるを得ない場合である。両者を分けて説明する。

5.1.1. 訓練事例追加の自動処理

機械学習の点からは、職業・産業コーディングは、コーディング作業が完了するたびに正解付き事例が蓄積されるという大きな利点がある。訓練事例のサイズが大きいほど精度も向上するため、正解付きの事例を訓練事例に追加することは有効である。

訓練事例として追加するには、まず、追加したい正解付き事例から、図 2-3 に示すような正解と素性から構成されるファイルを生成する必要がある。しかし、この作業はかなり面倒であるため、自動的に処理する機能を開発した。

Version7.4（未公開）からは、図 5-1 に示す操作画面により、希望する訓練事例の追加を行うことができる。操作方法は 4.3 節で説明した方法とほぼ同様であるが、最初に Update ボタンを押し、システムからの確認に OK ボタンを押して開始する点と、コード（複数可）にチェックした後に、Run ボタンではなく Update ボタンを押す点が異なる。



図 5-1 システム操作初期画面（Version7.4）

訓練事例に追加する正解付き事例の入力ファイル形式は、図 1-4 で説明した入力ファイルの右列に正解を付けたものである。正解を付ける列はコードの種類によって異なる。SSM 職業コード用訓練事例は G 列、SSM 産業業コード用訓練事例は H 列、ISCO 用訓練事例は I 列、ISIC 用訓練事例は K 列をそれぞれ使用する。使用しない列は、空欄のままにしておく。一度に複数の列を使用してもよく、4 列をすべて使用して、4 種類の訓練事例を一度に更新することも可能である。

例えば、既存の SSM 職業コード用訓練事例に、新規に正解付き事例を追加したい場合、まず図 5-2 のように、G 列に正解を入れた入力ファイルを用意する。次に、システムを立ち上げて図 5-1 を表示し、図 5-3 の手順を踏めば、図 5-2 が訓練事例に追加される。既存の訓練事例は一世代前まではバックアップが取られる。

A 列	B 列	C 列	D 列	E 列	F 列	G 列	H 列
11	10	9	保険会社の支店	保険会社の事務	11	559	
12	12	8	会社の売店	販売員	8	569	
62	9	7	夫の社会保険事務所	社会保険事務所の総務、経理	3	554	
289	9	10	農業	野菜を作っている	2	599	
465	9	1	訪問介護事業	訪問介護の経営、介護福祉士	4	801	

図 5-2 入力ファイルの例（SSM 職業コード用訓練事例追加の場合）

Update ボタンを押す → 訓練事例生成を確認するための OK ボタンを押す
 → Open ボタンを押して入力ファイルを指定 → 職業の SSM 欄にチェック
 → Update ボタンを押す

図 5-3 訓練事例追加の操作（SSM 職業コード用訓練事例の場合）

訓練事例追加処理の過程で、追加する正解付き事例に素性辞書にない単語が出現すると、どれも「200,000」なる番号が付けられるため、新出単語が増えるにつれ、システムの精度が低下してしまう可能性が高い。そこで、Version8.1（未公開）では、訓練事例を追加する際に素性辞書にない単語が出現すると素性辞書に登録することで、素性辞書の自動更新も同時に行うことができるようにした。1 回の訓練事例追加処理の最中に、再度その単語が出現した場合は、すでに登録されてあるため、該当する素性番号に変換される。

なお、訓練事例を追加する際に注意すべき点として、追加したい訓練事例が既存の訓練事例とコード付与のルールが大きく異なるような場合、訓練事例のサイズが拡大される効果は期待できず、むしろ精度が低下する可能性の問題がある。

5.1.2. 訓練事例全体の差し替えまたは新規追加

職業や産業のコードは、これまでは表 1-1 に示したものが用いられることが多かった。

しかし、今後は別のコード体系を利用する場合も出てくるであろう。もし、利用するコード体系と表 1-1 のコード体系との間の対応関係が簡単であれば、両者の対応表を作成しておき、現行のシステムで処理を行った後に対応表によるコード変換を行えばよい。しかし、これがむずかしい場合には、利用する訓練事例全体を変更する必要がある。

この場合は、新規の訓練事例が必要で、これを生成するためには、正解の付いた事例がある程度大量にあることが条件となる。もし正解の付いた事例が少ないまたは全くない場合は新たに正解を付ける必要があり、このための人手と時間がかかる。いったん正解付き事例が用意できれば、これを訓練事例として生成する作業は自動化されており、容易である（図 5-1 において、既存の訓練事例が空である特殊な場合と考えればよい）。

新規に生成された訓練事例全体を既存のものと変更する方法は、現時点では、差し替えを行うしかない。現行のシステムは操作の容易さを優先し、実行時に訓練事例を指定しなくてよいようにしたため、利用する訓練事例のファイル名を固定している。したがって、新たに利用する訓練事例のファイルに現行のファイル名を付けて利用する。

しかし、この方法は自由度に欠けるため、現在、システム操作用初期画面に複数種類の訓練事例を表示し、その中から選択できるように改良中である。システム操作用初期画面は、Version8.1（未公開）においても、職業の場合、「SSM」と「ISCO」のみが表示されるが（図 5-1 参照）、例えば、「SSM（95年版）」「SSM（15年版）」「ISCO-88」「ISCO-08」のように複数種類を表示し、チェックが付けられた訓練事例を生成することを目指している。

これは、自動コーディング処理においても同様のことがいえるため、次には自動コーディング処理における改善も予定している。

5.2. シソーラスの更新方法（ルールベース手法）

自由回答にこれまでは出現しなかった語が登場した場合、既存のシソーラスには存在しないため、この語がもつ情報が活かされない。機械学習における素性辞書と同様に、精度向上のためには、ルールベース手法におけるシソーラスも随時、更新していく必要がある。

本システムはこれを支援する機能を特には備えていないため、手作業で行う必要があるが、述語シソーラス（図 2-9 参照）と名詞シソーラス（図 2-10 参照）はいずれもテキスト形式のファイルであるため、更新作業は容易である。

本システムでは、訓練事例の場合と同様に、利用するシソーラスのファイル名を固定しているため、更新したシソーラスファイルに現行のシソーラスのファイル名を付けて差し替える必要がある。

5.3. ルール辞書の更新方法（ルールベース手法）

ルール辞書の更新は、次の 2 つの場合に必要なになる。

一つは、新しい語が出現したとき、シソーラス、特に述語シソーラスにおける新規の述語コードを追加し、これに対する SSM 職業／産業コードを決定するルールを追加する場合である。

もう一つは、新しく SSM 職業／産業コードが作成されたとき、これを決定するルールを既存の述語コードと対応させながら追加する場合である。

本システムはこれを支援する機能を特には備えていないため、手作業で行う必要があるが、職業ルール辞書 α（図 2-12 参照）と産業ルール辞書（図 2-13 参照）はいずれもテキスト形式のファイルであるため、更新作業は容易である。

本システムで利用するルール辞書もシソーラスの場合と同様にファイル名が固定されているため、更新したルール辞書ファイルに現行のルール辞書のファイル名を付けて差し替える必要がある。

6. システムの課題と対応

本節では、本システムの課題とその対応予定について述べる。

（１）精度の向上

精度（本報告書では正解率）はシステムの性能を示す重要な指標である。本システムの精度は十分に満足できる値ではなく、精度の向上は今後も最優先課題として継続する。

精度の向上はどの分類タスクにも共通の問題であるため、当初は汎用的な手法として、クラス所属確率を利用したアンサンブル学習（素性選択を違えて複数の分類器を生成し、事例ごとに、各分類器が予測した第１位のクラスに対するクラス所属確率を推定し、その中でもっとも大きな値をもつ分類器が予測したクラスに決定する）を提案した。

提案手法を職業・産業コーディングに適用した実験の結果、有効性を示したため、引き続き、公開データセットを用いた実験を行って汎用性を示すことと併行し、実際にこの手法を本システムに組み込む方法の検討を開始した。その結果、提案手法を自動化して組み込む作業は簡単ではない上に、もともと SVM は長時間を要するが、特にワークステーションと比較すると処理能力の劣るパソコン上で複数個の分類器を生成するのは長時間を要するため、これに見合うほどの有効性があるのかという点で、提案手法を組み込むことは中止せざるを得なかった。そこで、本タスクに絞った精度の向上を目指すことにした。

調査の結果、ルールベース手法による結果の正解・不正解が、システムの精度に大きな影響を与えることがわかったため、現在、既存のシソーラスやルール辞書の改善、素性辞書の更新、訓練事例の見直しを行っている。

精度の向上と関連するが、コードの作業量軽減という点では、確信度 A が付与された事例の正解率や再現率も重要である。正解率については一応の目標を達成しているが、上記の改善作業は、確信度 A の再現率の向上にもつながるものと考えている。

（２）コード変更への対応

現在用いられているコードは、システム開発時のものとやや異なっている。例えば、SSM 職業コードの場合、当初は『SSM 産業分類・職業分類（95 年版）』に記載されたコード（500 番台と 600 番台）であったが、現在は資料(1)に示すとおり、700 番台が追加され、さらに最近では 800 番台のコードが追加されている。産業コードも同様で、一部のコードが分化している。

本システムが 2015SSM 調査に利用されたのを機に、コードの見直しを行ったが、十分であったとはいえない。今後、これを徹底させ、シソーラスやルール辞書の更新を行っていく予定である。

（３）コード体系変更への対応

同一コード体系内での修正は上記（２）で対応できるが、これまでとは異なるコード体系が用いられる場合は、訓練事例全体を変更する必要がある（5.1.2 節を参照のこと）。

これまで用いられてきたコードと新規のコード間の対応関係が簡単なる場合は問題が少ないが、そうではない場合には、新規の訓練事例が必要なため、これがない場合には、訓練事例を生成するための正解付きの事例を用意する必要がある。

（４）利用方法の再考

現在は、システムの利用方法を、「システム自体を公開し、利用者自身がシステムをダウンロードして実行する方法」とはしていない。昨今の ICT に関する知識の高まりをみると、利用者がソフトウェア環境を整えることはそれほど困難ではないかもしれないと考え、参考のために 4 節で説明を行った。しかし、利用者個々のコンピュータ環境（OS のバージョンの違い）まで含めたサポート体制は実現不可能であり、今後の課題としたい。

（５）データの質の向上

この問題はどの分野においても重要な課題であるが、職業・産業コーディングの場合は、特にコンピュータに入力された自由回答の情報がコードの決定に大きく影響するため、過不足のない内容をもつ自由回答を収集することが必要になる。自由回答にコードを決定するための情報が存在しない場合には、正確なコーディングができないだけでなく、作業効率が低下し、コードもストレスを感じるであろう。

問題は、何が必要な情報なのかを、コードの内容を熟知していない回答者や調査員にはわからないことである。そこで、最近では、よくあるケースについては、注意事項をあらかじめ調査票に記載したり、調査員へのインストラクション時に喚起したりすることも行われている。例えば、自由回答に「営業」としか書かれていない場合、「557」（営業事務）なのか「573」（外交員（保険、不動産を除く））なのか判断しにくいいため、「営業」なる回答には、続けて、内勤か外勤かも尋ねるように指示されることが多くなった。

このようにして、調査の現場で回答者自身から必要な情報を収集できれば、アフターコーディングが正確かつスムーズにいく可能性が高まる。しかし、すべてのコードに対してこのような指示を与えるのは困難であり、調査員の負担も大幅に増えることになる。職業・産業コードをすべて選択肢として提示する方が、むしろ双方とも負担が少ないのではないかという議論になりかねない。

そこで、「調査現場にコンピュータを持ち込んで回答を入力し、コードを決定するために不足している情報があれば、その場で追加質問をして情報収集を行うシステム」（図 6-1 参照）の構築を計画している。

このようなシステムが開発できれば、調査現場でデータの電子化が行えるため、副次的な効果として、職業や産業に関するデータについては、入力作業が不要となる。また、こ

れまでのように、調査員が調査票に回答を書き込む場合は、漢字と平仮名またはカタカナが混在する語の問題や誤字・脱字の問題が避けられないが、直接、コンピュータに入力するのであれば、漢字変換が行えるため、これらの問題が減少することが期待できる。融通のきく人間と違い、コンピュータによる自動化処理を行う本システムにとっては、この改善は形態素解析の成功率を向上させる効果があるため、精度の向上に大きく貢献する。

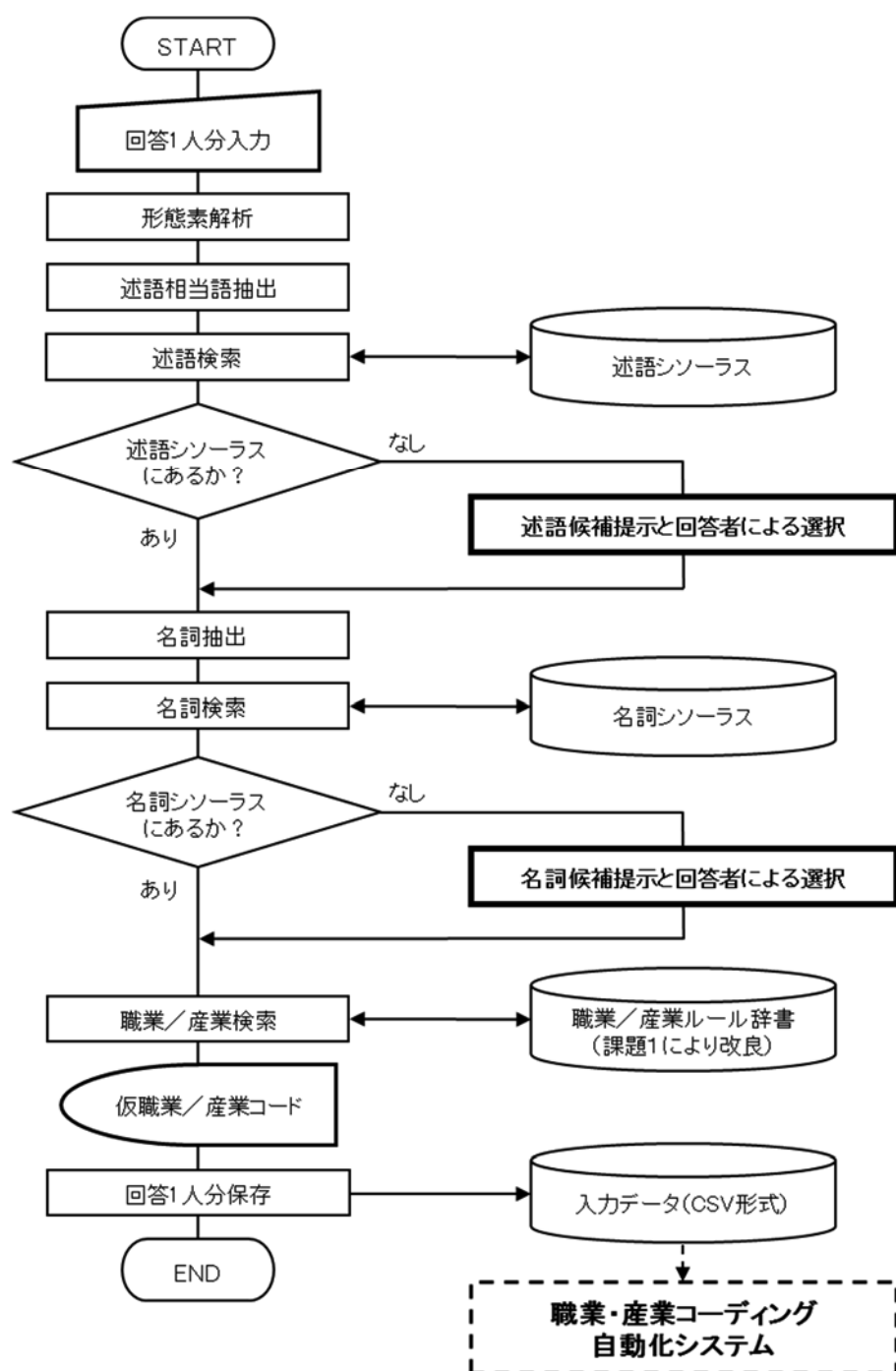


図 6-1 職業・産業コーディングにおけるデータの質の向上（案）

7. 自由回答一般への拡張可能性

本システムは、見方を変えれば、3種類の選択回答と2種類の自由回答からなるデータを1個のクラスに分類するタスクであるといえる。そこで、これを自然に拡張し、「2種類以内の自由回答と3種類以内の選択回答からなるデータ」を1個のクラスに分類するシステムとできるのではないかと考え、開発を進めている。まだ構想段階ではあるが、以下で概要を説明する。

この拡張システムは、入力データに自由回答や選択回答から総合的に分類をする場合を想定しているが、自由回答だけで選択回答がなくてもよい（選択回答だけのデータに対しても処理は行う）。

入力ファイルの形式は、図 7-1 に示すように、通し番号の後に、選択回答を3種類、自由回答を2種類続けて入力する。データにより、すべての回答欄が必要でない場合には、選択回答、自由回答のそれぞれで左列に詰め、使用しない欄は空白のままにしておく。例えば、データが1種類の自由回答だけの場合は、A列とE列のみを使用することになる。

<i>A 列</i>	<i>B 列</i>	<i>C 列</i>	<i>D 列</i>	<i>E 列</i>	<i>F 列</i>
通し番号	選択回答 1	選択回答 2	選択回答 3	自由回答 1	自由回答 2

図 7-1 拡張システムにおける入力ファイルの形式（案）

結果ファイルは、図 1-5 に示した形式と同様のものを想定している。

この拡張システムも SVM を適用するため、訓練事例が必要である。訓練事例の生成は、正解付きの事例が蓄積されていない場合にはコストがかかるため、正解付きの事例がない1回限りの調査には不向きである。しかし、訓練事例のサイズが大きくなるほど精度が向上するため、繰り返し実施する調査には有効であると思われる。

拡張システムの操作用初期画面を図 7-2 に示す。Open ボタンを押して、入力ファイル（上段）と訓練事例（下段）を指定した後、Run ボタンを押して実行を開始する。

今後、Update ボタンによる訓練事例の追加機能も追加する予定である。

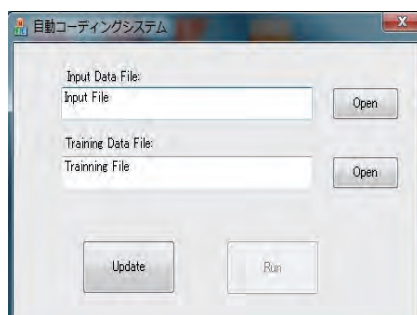


図 7-2 拡張システムの操作用初期画面（案）

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。

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。

参考文献（資料編掲載以外）

- ・1995 年 SSM 調査研究会, 2006, 『SSM 産業分類・職業分類（95 年版）修正版』1995 年 SSM 調査研究会.
- ・1995 年 SSM 調査研究会（編）, 1996, 『1995SSM 調査コード・ブック』1995 年 SSM 調査研究会.
- ・2005 年社会階層と社会移動調査研究会, 2007, 『2005 年 SSM 日本調査コード・ブック』2005 年社会階層と社会移動調査研究会.
- ・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.
- ・三輪哲（編）, 2011, 『SSM 職業分類・産業分類の改定に向けて』科学研究費補助金基盤研究 A「現代日本の階層状況の解明—ミクロ・マクロ連結からのアプローチ」研究成果報告書別冊.
- ・大阪商業大学比較地域研究所・東京大学社会科学研究所, 2005, 『日本版 General Social Surveys 基礎集計表・コードブック JGSS-2003』大阪商業大学比較地域研究所.
- ・労働政策研究・研修機構（編）, 2011, 『第 4 回改訂 厚生労働省編職業分類 職業名索引』労働政策研究・研修機構.
- ・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.
- ・田辺俊介・相澤真一, 2008, 『東京大学社会科学研究所. 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』東京大学社会科学研究所.
- ・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, 共同調査と共同研究「自動コーディング（職業・産業）」(<http://csrda.iss.u-tokyo.ac.jp/joint/autocode/>) .

資料編

問2【回答票1】あなたの現在のお仕事についてうかがいます。複数の仕事をお持ちの場合は、主な仕事についてうかがいます。〔以下、aからfまで順に聞く〕

a 従業上の地位	あなたのお仕事は大きく分けてこの中のどれにあたりますか。 〔回答票から選んでもらい、該当する番号に○をつける〕	1 (ア) 経営者、役員 2 (イ) 常時雇用されている一般従業者 3 (ウ) 臨時雇用・パート・アルバイト 4 (エ) 派遣社員 5 (オ) 契約社員、嘱託	6 (カ) 自営業主、自由業者 7 (キ) 家族従業者 8 (ク) 内職 → d、e、fのみ聞く 9 (ケ) 無職：仕事を探している → 問6へ 10 (コ) 無職：仕事を探していない → 問6へ 11 (サ) 学生 → 問6へ 99 わからない	②⑥ ②⑦	
b 従業先の事業内容	あなたの勤め先は、どのような事業をいとなんでいますか。〔派遣社員は派遣会社を勤め先とする〕	〔野菜の販売、自動車の製造、薬品の卸売、衣服の小売、旅館経営のように具体的に記入すること〕 99 わからない			②⑧ ②⑨
c 従業員数	従業員（働いている人）は、会社全体で何人ぐらいですか。〔家族従業者、パート・アルバイトも含む〕	1 (ア) 1人 2 (イ) 2～4人 3 (ウ) 5～9人 4 (エ) 10～29人	5 (オ) 30～99人 6 (カ) 100～299人 7 (キ) 300～499人 8 (ク) 500～999人	9 (ケ) 1000人以上 10 (コ) 官公庁 99 わからない	③⑩ ③⑪
d 本人の仕事の内容	あなたは職場でどのような仕事をしていますか。具体的な仕事の内容を教えてください。	〔小学校教員、プラスチック製品（おもちゃ）の製造、スーパーのレジ係、銀行の窓口の仕事、高齢者家庭で身の回りの世話・介護、など仕事がわかるように具体的に記入すること〕 999 わからない			③② ③③ ③④
e 役職名	何かの役職についていますか。 〔該当するものに○をつける〕	1 (ア) 役職なし 2 (イ) 監督、職長、班長、組長 3 (ウ) 係長、係長相当職 4 (エ) 課長、課長相当職 5 (オ) 部長、部長相当職	6 (カ) 社長、重役、役員、理事 7 (キ) その他（具体的に） 〔 〕 9 わからない	③⑤	
f 労働時間	このお仕事をふだん1日何時間、週何日していますか。（または月何日していますか。）残業の時間も含めてください。	1日 <input type="text"/> 時間 99 わからない 週 <input type="text"/> 日 または 月 <input type="text"/> 日 99 わからない 隔週勤務など、特殊なケースについては以下に記入しておくこと 〔 〕			③⑥③⑦ ③⑧ ③⑨④⑩

4. 産業コード

SSM 産業分類（95 年版）

* JGSS 用追加コード含む

2005 年コード	1995 年コード
010 農業	01 農業
020 林業	02 林業
030 漁業	03 漁業
040 鉱業	04 鉱業
050 建設業	05 建設業
060 製造業	06 製造業
070 電気・ガス・熱供給・水道業	07 電気・ガス・熱供給・水道業
081 運輸業	08 運輸業
082 旅行業	
091 卸売業	09 卸売・小売業, 飲食店
092 小売業	
093 飲食店	
100 金融・保険業	10 金融・保険業
110 不動産業	11 不動産業
121 新聞・放送・出版業, 映画制作業	12 新聞・放送・出版業, 広告業, 映画制作業
122 広告業	
131 情報・通信サービス業	13 情報・通信サービス業
132 郵貯・簡保	
140 医療・福祉サービス業	14 医療・福祉サービス業
150 教育・研究サービス業	15 教育・研究サービス業
160 法律・会計サービス業	16 法律・会計サービス業
171 その他のサービス業	17 その他のサービス業
172 学習塾・教養技能・スポーツ施設	
180 公務	18 公務
190 分類不能の産業	19 分類不能の産業
980 非該当	98 非該当
990 不明, 無回答	99 不明, 無回答

SSM 産業コード（95 年版）は平成 7 年国勢調査産業分類にもとづいて作成されたものである。国勢調査分類との対応については 1995 年 SSM 調査研究会（編）『1995 年 SSM 調査コード・ブック』（p.91～）を参照のこと。

なお、2005 年 SSM 調査では 95 年版の産業コードに若干の修正を加えたものを使用している。修正の内容については、本コード・ブック「6. SSM 産業・職業のコーディング・ガイド」を参照。

5. 職業コード

SSM 職業分類（95 年版）

* JGSS 用追加コード含む

1. 専門的・技術的職業従事者

501	自然科学系研究者
502	人文科学系研究者
503	機械・電気・化学技術者
504	建築・土木技術者
505	農林技術者
506	情報処理技術者
507	その他の技師・技術者
508	医師
509	歯科医師
510	薬剤師
511	助産婦
512	保健婦
513	栄養士
514	看護婦，看護師
515	あん摩・はり・きゅう師，柔道整復師
516	その他の保健医療従事者
517	裁判官，検察官，弁護士
518	その他の法務従事者
519	公認会計士，税理士
520	幼稚園教員
521	小学校教員
522	中学校教員
523	高等学校教員
524	大学教員
525	盲・ろう・養護学校教員
526	その他の教員
527	宗教家
528	文芸家，著述家
529	記者，編集者
530	彫刻家，画家，工芸美術家
531	デザイナー
532	写真家，カメラマン
533	音楽家（個人に教授するものを除く）
534	俳優，舞踊家，演芸家 （個人に教授するものを除く）
535	職業スポーツ家 （個人に教授するものを除く）
536	獣医師

537	保母，保父
538	社会福祉事業専門職員
539	個人教師
540	不動産鑑定士
541	経営コンサルタント
542	アナウンサー（ラジオ・テレビ）
543	図書館司書
544	その他の専門的・技術的職業従事者

2. 管理的職業従事者

545	管理的公務員
546	国会議員
547	地方議員
548	会社役員
549	その他の法人・団体の役員
550	会社・団体等の管理職員
551	駅長，区長
552	郵便局長，電報・電話局長
553	その他の管理的職業従事者

3. 事務的職業従事者

554	総務・企画事務員
555	受付・案内事務員
556	出荷・受荷事務員
557	営業・販売事務員
558	その他の一般事務員
559	会計事務員
560	郵便・通信事務員
561	集金人
562	その他の外勤事務従事者
563	運輸事務員
564	速記者，タイピスト，キーパンチャー
565	電子計算機等操作員

4. 販売的職業従事者

566	小売店主
567	卸売店主
568	飲食店主
569	販売店員

- 570 行商人，呼売人，露天商人
- 571 再生資源卸売人・回収人
- 572 商品仲立人
- 573 外交員（保険，不動産を除く）
- 574 保険代理人・外交員
- 575 不動産仲介人・売買人
- 576 質屋店主・店員
- 577 その他の販売類似職業従事者

5. サービス的職業従事者

- 578 女中，家政婦，家事サービス職業従事者
- 579 理容師，美容師
- 580 クリーニング職，洗張職
- 581 料理人
- 582 パーティンダー
- 583 給仕係
- 584 スチュワーデス，スチュワード
- 585 接客社交係
- 586 娯楽場等の接客員
- 587 旅行・観光案内人
- 588 その他の個人サービス職業従事者
- 589 旅館・貸席等の主人・番頭，ホテル支配人
- 590 下宿・アパートの管理人，舎監，寮母
- 591 ファッションモデル
- 592 その他のサービス職業従事者

6. 保安的職業従事者

- 593 自衛官
- 594 警察官，海上保安官，鉄道公安員
- 595 消防員
- 596 看守，守衛，監視員
- 597 その他の保安職業従事者
- 598 旧職業軍人

7. 農林的職業従事者

- 599 農耕・養蚕作業者
- 600 植木職，造園師
- 601 畜産作業者
- 602 林業作業者
- 603 その他の農林業作業者
- 604 漁業作業者
- 605 漁船の船長・航海士・機関長・機関士

8. 運輸・通信従事者

- 606 電車・機関車運転士
- 607 自動車運転者

- 608 船長・航海士（漁船を除く），水先人
- 609 船舶機関長・機関士（漁船を除く）
- 610 航空機操縦士，航空士，航空機関士
- 611 車掌
- 612 鉄道員
- 613 船員
- 614 その他の運輸従事者
- 615 無線通信士，無線技術士
- 616 有線通信士
- 617 電話交換手
- 618 郵便・電報外務員
- 619 その他の通信従事者

9. 採掘作業者

- 620 採鉱員，採炭員
- 621 石切出作業者
- 622 その他の採掘作業者

10. 窯業・土石製品・金属材料・化学製品製造作業者

- 623 陶磁器工，絵付作業者
- 624 石工
- 625 ガラス・セメント製品製造作業者
- 626 その他の窯業・土石製品製造作業者
- 627 製鉄工，製鋼工，精錬工
- 628 鋳物工，鍛造工，金属材料製造作業者
- 629 化学製品製造作業者

11. 金属製品・機械製造作業者

- 630 金属工作機械工，めっき工，金属加工作業者
- 631 鉄工，板金工
- 632 金属溶接工
- 633 一般機械器具組立工・修理工
- 634 電気機械器具組立工・修理工
- 635 自動車組立工・整備工
- 636 鉄道車両組立工・修理工
- 637 船舶ぎ装工（他に分類されない）
- 638 航空機組立工・整備工
- 639 自転車組立工・修理工
- 640 その他の輸送機械組立・修理作業者
- 641 時計組立工・修理工
- 642 光学機械・精密機械器具組立工・修理工

12. その他の製品製造作業者

- 643 精穀工，製粉工
- 644 パン・菓子・めん類・豆腐製造工

- 645 味噌・醤油・缶詰食品・乳製品製造工,
飲食料品製造作業
- 646 たばこ製造工
- 647 酒類製造工
- 648 製糸業者
- 649 織布工, 紡織業者
- 650 漂白工, 染色工
- 651 洋服・和服仕立職
- 652 縫製工, 裁断工
- 653 製材工, 木工
- 654 指物職, 家具職, 建具職
- 655 船大工
- 656 おけ職, 木・竹・草・つる製品製造
作業
- 657 製紙工, 紙器製造工, パルプ・紙・
紙製品製造作業
- 658 印刷・製本業者
- 659 ゴム・プラスチック製品製造業者
- 660 くつ製造工・修理工, かわ・かわ製品
製造業者
- 661 塗装工, 画工, 看板工
- 662 漆塗師, まき絵師
- 663 表具師, 内張工
- 664 和がさ・ちょうちん・うちわ職
- 665 貴金属・宝石・甲・角等細工工
- 666 印判師
- 667 洋傘組立工
- 668 かばん・袋物製造工
- 669 がん具製造工
- 670 製図工, 現図工
- 671 映写技士
- 672 その他の技能工・生産工程作業

13. 定置機関運転・建設機械運転・電気作業

- 673 汽かん士, 汽かん火夫
- 674 起重機・建設機械運転作業
- 675 その他の定置機関運転作業
- 676 発電員, 変電員
- 677 電気工事・電話工事作業

14. 建設作業

- 678 土木・建築請負師
- 679 左官, とび職
- 680 れんが積工, 配管工
- 681 畳職
- 682 土工, 道路工夫

- 683 鉄道線路工夫
- 684 現場監督, その他の建設作業

15. 労務作業

- 685 倉庫夫, 仲仕
- 686 運搬労務者
- 687 清掃員
- 688 その他の労務作業

16. その他

- 689 分類不能の職業
- 690 旧地主
- 691 名目上の役員
- 701 スーパー等のレジスター係員,
キャッシャー
- 702 大工
- 703 教員: 小学校・中学校・高校などが明記
されていない場合
- 704 製品製造作業: (特に父職で) 作って
いる製品が明記されていない場合
- 705 会社員: (特に父職で) 記入が「会社員」
とあった場合
- 706 「607 自動車運転者」「686 運搬労務者」
のいずれにも該当しないもの
- 707 自営業: (特に父職で) 記入が「自営業」
とあった場合
- 998 非該当
- 999 不明, 無回答

6. 産業・職業のコーディング・ガイド

(a) 従業上の地位

- ・ 「経営者、役員」と「自営業主、自由業者」とは内容的には区別しないで、回答のままとする。
- ・ 勤労働員・学徒動員は「兵役」でコードする。
- ・ 兼業農家について

夫の職業が農業ではなくて、妻が農業をやっていることがある。その時、妻は、「6 家族従業者」とコードされるが、夫は必ずしも「自営業主」ではない。

(b) 従業先の事業内容（産業）

(1) 産業コードの変更と追加コード

SSM95 産業・職業分類コード（『SSM 産業分類・職業分類（95 年版）』参照）の各産業コードの下 1 桁にゼロ（0）を付け加え、全体のコードを 3 桁とする変更を施した。たとえば、「010 農業」「020 林業」「030 漁業」「040 鉱業」などとなる。

さらに、以下のような新コードを追加した。

95 年コード	05 年コード	変更内容
08 運輸業	081 運輸業	運輸業の中の旅行業（旅行代理店など）は日本標準産業分類（第 11 回改訂）「Q その他のサービス業」に分類される。
	082 運輸業（旅行業）	
09 卸売・小売業、飲食店	091 卸売業	「09 卸売・小売業、飲食店」は、細かい分析が可能となるよう 3 つに分類する。
	092 小売業	
	093 飲食店	
12 新聞・放送・出版業、広告業、映画制作業	121 新聞・放送・出版業、映画制作業	広告業は、日本標準産業分類（第 11 回改訂）にともない、「Q その他のサービス業」に分類される。
	122 広告業	
13 情報・通信サービス業	131 情報・通信サービス業	郵貯、簡保は、日本標準産業分類（第 11 回改訂）の「K 金融・保険業」に分類される。
	132 郵貯・簡保	
17 その他のサービス業	171 その他のサービス業	「17 その他のサービス業」から、「趣味・工芸・学習などの個人教授、スポーツ施設」を区別する。
	172 学習塾・教養技能・スポーツ施設	

(2) 注意を要する産業分類例

「造園業」「植木職」	010 農業	その他のサービス業ではない
「電気（設備）工事」	050 建設業	電気業ではない
「印刷」「製本」	060 製造業	出版業ではない
「道路公団」	081 運輸業	公務ではない
「豆腐屋、菓子屋など」	092 小売業	製造業でない
「弁当屋」	092 小売業	製造業ではない
「ガソリンスタンド」	092 小売業	その他のサービス業ではない

「質屋」	100	金融・保険業	その他のサービス業ではない
「レコード・CD 製作」	121	新聞・放送・出版等	製造業ではない
「保育所」「託児所」	140	医療・福祉サービス業	その他のサービス業ではない
「学校給食」	150	教育・研究サービス業	その他のサービス業ではない
「公民館」	150	教育・研究サービス業	その他のサービス業ではない
「大学病院」	150	教育・研究サービス業	医療・福祉サービス業ではない
「人材派遣」	171	その他のサービス業	
「設計事務所」	171	その他のサービス業	
「職安」	180	公務	
「教育委員会」	180	公務	

(3) その他

・ 官公署の扱い

本来の立法・司法・行政事務を行う場合、権力的な業務を行う場合は「18 公務」に分類されるが、それ以外は、その内容によってそれぞれの産業に分類する。

ただし、従業員数はすべて「10 官公庁」である。

○税務署、消防署、警察署 → 「18 公務」

○自衛隊、進駐軍（昭和 26 年以前）→ 「18 公務」

×駐在軍（昭和 27 年以降）→ 「17 その他のサービス業」

×陸軍・海軍の被服工廠（第二次世界大戦以前）→ 「06 製造業」

・ 国鉄、電電公社、専売公社、郵便局・郵政公社

国鉄、電電公社、専売公社はいずれも官公庁ではなく、事業員数は「9 1000 人以上」となる。また、郵便局および郵政公社（調査時）は官公庁とする。

なお、国鉄と JR は別企業とみなすが、電電公社と NTT（日本電信電話）、専売公社と JT（日本たばこ）は同一企業とみなす。

・ 製造小売業

製造した商品をもその場所で個人または家庭用消費者に販売する洋服店、菓子店、パン屋、豆腐屋、家具屋、建具屋、畳屋などは、製造業としないで「09 卸売・小売業 飲食店」に分類される。（ただし、職業については一般に「パン職人」や「製造工」のように製造の方を優先してコードする。）

(c) 従業員数（規模）

「官公庁」とは、各省庁およびその出先機関、県市区町村役場、国公立学校などを指す。郵便局は官公庁に含むが、それ以外の公団・公社・独立行政法人は含まない。

(d) 仕事の内容（職業）

(1) 職業の追加コード

SSM95職業小分類188の分類カテゴリーに、以下のような新カテゴリー追加した。

701 レジ・キャッシャー	「559 会計事務員」の中の「スーパーなどのレジスター係員・キャッシャー」を別カテゴリーとしてとして独立させる。
702 大工	「679 大工、左官、とび職」の「大工」を別カテゴリーとしてとして独立させる。
703 教員	小学校、中学校、高校などの区別のないときに用いる。
704 製品製造作業	何の製品かわからないが工場で製造しているときに用いる。
705 会社員	(特に父職)会社員としか記入のないときに用いる。
706 宅配便の配達	「607 自動車運転者」「686 運搬労働者」のいずれにも該当しないため、独立したカテゴリーとする。
707 自営業	(特に父職)自営としか記入のないときに用いる。

(2) 「管理職」のコードについて

(以下は原則であり、仕事の内容が管理的としか言えないときは、管理的職業にコードする。ただし、ここでいう「管理」とは部下など人の管理であって、在庫管理や物品管理などは含まないことに注意。)

- ・ 従業上の地位が「1 経営者・役員」もしくは「6 自営業主・自由業者」の場合
 - 規模 5 人未満 ・ 必ず、管理的職業以外の仕事の内容でコードする。
 - 規模 30 人未満 ・ 管理的職業以外の仕事の内容を優先してコードする。
 - 規模 30 人以上 ・ 原則としていずれか該当する管理的職業でコード（「548 会社役員」「549 その他の法人・団体の役員」など）するが、それ以外の仕事の内容が書いてあれば、それに従ってコードする。
- ・ 従業上の地位が「2 常時雇用されている一般従業者」「3 臨時雇用・パート・アルバイト」「4 派遣社員」「5 契約社員、嘱託」「7 家族従業者」の場合
 - 役職が「4 課長」以上のとき
 - 規模 5 人未満 ・ 必ず、管理的職業以外の仕事の内容でコードする。
 - 規模 30 人未満 ・ 管理的職業以外の仕事の内容を優先してコードする。
 - 規模 30 人以上 ・ 原則としていずれか該当する管理的職業でコード（「550 会社・団体等の管理職員」など）するが、それ以外の仕事の内容が書いてあれば、それに従ってコードする。
 - 役職が「3 係長」以下のとき
 - 必ず、管理的職業以外の仕事の内容でコードする。
- ・ 専門的管理職（設計技師長、病院長、学校長など）は、「専門」の方を優先する。

(3) 「土木・建設・建築の仕事」、「土建業」、そして「現場監督」のコードについて

- ・ 従業上の地位が「1 経営者・役員」「6 自営業主・自由業者」の場合
 - 規模 30 人未満 ・ 「678 土木・建築請負師」
 - 規模 30 人以上 ・ 「548 会社役員」
- ・ 従業上の地位が「2 常時雇用されている一般従業者」「3 臨時雇用・パート・アルバイト」

イト」「4 派遣社員」「5 契約社員、嘱託」の場合

- 役職が「なし」のとき…
 - 「682 土工」
- 役職が「2 監督・班長」「3 係長」「4 課長」のとき…
 - 「684 現場監督」
- 役職が「5 部長」のとき…
 - 規模 30 人未満 「684 現場監督」
 - 規模 30 人以上 「550 会社の管理職員」

- ・ 「(自由記述が) 現場監督」
 - 自営業主(「6 自営業主・自由業者」)の場合は「678 土木・建築請負師」
 - 多少でも仕事についても記述があれば「679 左官・とび職」、「702 大工」などに分類する。
 - 非自営で、大卒の場合 「504 建築・土木技術者」
 - 上記以外の場合 「684 現場監督・その他の建築作業者」

(4) 注意を要する職業分類例

(事務、営業関連の職業)

医療事務	558	一般事務	
「コールセンター」	617	電話交換手	ただし販売勧誘も行う場合には「557 営業・販売事務員」
「電話での販売」「テレフォン・アポインター」	557	営業・販売事務員	
「(コンビニ)店長」	566	小売店主	従業上地位が「経営者」「自営業主」、または一般従業者で役職「係長」以下
	550	会社・団体の管理職員	一般従業者で役職「課長」以上
	569	販売店員	従業上地位が「アルバイト、派遣、契約」
「(レストラン)店長」	568	飲食店主	従業上地位が「経営者」「自営業主」、または一般従業者で役職「係長」以下
	550	会社・団体の管理職員	一般従業者で役職「課長」以上
	583	給仕係	従業上地位が「アルバイト、派遣、契約」

(外勤関連の職業)

「(世論)調査員」	562	その他の外勤事務	
「保険調査員」	557	営業・販売事務員	
「バイク便」	706	宅配便配達	
「ピザ宅配」	686	運搬労務者	牛乳配達員と同じ

(製造関係)

「自動車部品製造」	633	一般機械器具組立工	自動車組立工、整備工ではない
-----------	-----	-----------	----------------

(サービス関連)

「スーパーで資材の搬入」	685	倉庫夫	
--------------	-----	-----	--

「(スーパーの精肉・鮮魚部)肉・魚の解体」	672	その他の技能工	
「豆腐の製造・販売」	644	パン・菓子・麺類・豆腐製造工	ただし販売のみは「566 小売店主」「569 販売店員」
「お弁当製造・販売」	645	飲食料品製造作業	ただし販売のみは「566 小売店主」「569 販売店員」
「(居酒屋の)店員」	583	給仕係	×582 バーテンダー ×569 販売店員
「皿洗い」「洗い場」	583	給仕係	
「自動車教習所指導員」	607	自動車運転者	
「テニス公認指導員」	539	個人教師	
「通信指導添削」	539	個人教師	
「レンタルビデオ店員」「レンタカー受付」	592	その他のサービス業	

(福祉関連)

「ケアマネジャー」	538	社会福祉事業専門職員	
「介護(福祉)士」	516	その他の保険医療従事者	資格がある場合
「介護、ヘルパー、世話」	578	女中、家政婦、家事サービス職業従事者	資格がない場合
「(社会)福祉士」	538	社会福祉事業専門職員	資格がある場合
「児童福祉士」	538	社会福祉事業専門職員	
「学童保育・学童クラブの指導員」	590	下宿・アパートの管理人、舎監、寮母	
「用務員」	688	その他の労務作業	
「旅館の女中・仲居」	583	給仕係	

(コンピューター関連)

「システムエンジニア」「プログラマー」「HP作成」「ウェブデザイナー」	506	情報処理技術者	
「HP管理」	565	電子計算機等操作員	
「CADオペレーター」	670	製図工・現図工	

(建築関連)

「建築士」	504	建築・土木技術者	
「設計」「建築設計」	504	建築・土木技術者	大卒以上
	670	製図工・現図工	その他
「室内装飾」	684	現場監督、その他の建設作業	
「壁紙はり、ふすまはり」	663	表具師	

・その他

議員 → 従業上の地位は「1 経営者・役員」とし、規模を「10 官公庁」とする。

(e) 役職

原則として、下のコードに対応する役職名がある場合には、そのコードを採用する。以下は役職名がプリ・コードの名称と異なるときの対応のさせ方。

- | | |
|----------------|--|
| 1. 役職なし | ○自営業主、家族従業者 |
| 2. 監督、職長、班長、組長 | ○主任
○現場監督
○巡査部長、軍人の下士官以下の役職 |
| 3. 係長、係長相当職 | ○課長補佐
○警部、警部補
○鉄道助役
○婦長
×主事（公務員） → 1 |
| 4. 課長、課長相当職 | ○調査役、参事、支店長、店長、営業所長、事務長
（※ただし規模が小さい場合は「3 係長」とした）
○支店長代理、郵便局長代理、課長代理
○駅長、郵便局長
ただし、大きな駅・郵便局の場合→5
○小・中・高校の校長・教頭
○園長
○警視、軍人の尉官以上
○30 人くらいの工場長、所長
○小さな船の船長
×大企業の支店長→5 |
| 5. 部長、部長相当職 | ○局長、次長
○大きな船の船長・機関長
○市町村の助役・収入役
○都道府県会議員、市町村会議員
○病院長
○1000 人くらいの工場長、所長
○自治体監査委員 |
| 6. 社長、重役、役員、理事 | ○会長、専務、常務、取締役、監査役
○政府・国会・裁判所高官、国会議員
○都道府県知事、市町村長
○宗教団体の役員 |

(f) 管理的職業コードの注意点

1985 年以前のデータのコーディングでは、管理的職業であるか否かの一つの判定基準として、規模が「5 人以上」かそうでないかをめやすにしていたが、2005 年調査では 1995 年調査と同様に、規模の基準を「30 人以上」にやや厳しく設定した。ただし、この規模の基準はどちらの場合も「原則として」であって、実際には他の情報を考慮しながら総合的にコードしてある。

また 2005 年調査では、規模 30 人以上で、本来は管理的職業であっても、職業コードとしては管理的とされていないケースが少なくない。これは、なるべく多様な情報を残すための処置であるが、過去の SSM コードとの比較のためにはこのままでは不適切であり、分析においては必要に応じて次のようなコード変換を行う必要がある。

7. 國際標準産業・職業分類コード

7.1 國際標準産業分類 (ISIC)

- A Agriculture, hunting and forestry
 - 1 Agriculture, hunting and related service activities
 - 2 Forestry, logging and related service activities
- B Fishing
 - 5 Fishing, operation of fish hatcheries and fish farms; service activities incidental to fishing
- C Mining and quarrying
 - 10 Mining of coal and lignite; extraction of peat
 - 11 Extraction of crude petroleum and natural gas; service activities incidental to oil and gas extraction excluding surveying
 - 12 Mining of uranium and thorium ores
 - 13 Mining of metal ores
 - 14 Other mining and quarrying
- D Manufacturing
 - 15 Manufacture of food products and beverages
 - 16 Manufacture of tobacco products
 - 17 Manufacture of textiles
 - 18 Manufacture of wearing apparel; dressing and dyeing of fur
 - 19 Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, harness and footwear
 - 20 Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
 - 21 Manufacture of paper and paper products
 - 22 Publishing, printing and reproduction of recorded media
 - 23 Manufacture of coke, refined petroleum products and nuclear fuel
 - 24 Manufacture of chemicals and chemical products
 - 25 Manufacture of rubber and plastics products
 - 26 Manufacture of other non-metallic mineral products
 - 27 Manufacture of basic metals
 - 28 Manufacture of fabricated metal products, except machinery and equipment
 - 29 Manufacture of machinery and equipment n.e.c.
 - 30 Manufacture of office, accounting and computing machinery
 - 31 Manufacture of electrical machinery and apparatus n.e.c.
 - 32 Manufacture of radio, television and communication equipment and apparatus
 - 33 Manufacture of medical, precision and optical instruments, watches and clocks
 - 34 Manufacture of motor vehicles, trailers and semi-trailers
 - 35 Manufacture of other transport equipment
 - 36 Manufacture of furniture; manufacturing n.e.c.
 - 37 Recycling
- E Electricity, gas and water supply
 - 40 Electricity, gas, steam and hot water supply
 - 41 Collection, purification and distribution of water
- F Construction
 - 45 Construction
- G Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods
 - 50 Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel
 - 51 Wholesale trade and commission trade, except of motor vehicles and motorcycles
 - 52 Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods

- H Hotels and restaurants
 - 55 Hotels and restaurants
- I Transport, storage and communications
 - 60 Land transport; transport via pipelines
 - 61 Water transport
 - 62 Air transport
 - 63 Supporting and auxiliary transport activities; activities of travel agencies
 - 64 Post and telecommunications
- J Financial intermediation
 - 65 Financial intermediation, except insurance and pension funding
 - 66 Insurance and pension funding, except compulsory social security
 - 67 Activities auxiliary to financial intermediation
- K Real estate, renting and business activities
 - 70 Real estate activities
 - 71 Renting of machinery and equipment without operator and of personal and household goods
 - 72 Computer and related activities
 - 73 Research and development
 - 74 Other business activities
- L Public administration and defence; compulsory social security
 - 75 Public administration and defense; compulsory social security
- M Education
 - 80 Education
- N Health and social work
 - 85 Health and social work
- O Other community, social and personal service activities
 - 90 Sewage and refuse disposal, sanitation and similar activities
 - 91 Activities of membership organizations n.e.c.
 - 92 Recreational, cultural and sporting activities
 - 93 Other service activities
- P Activities of private households as employers and undifferentiated production activities of private households
 - 95 Private households with employed persons
- Q Extraterritorial organizations and bodies
 - 99 Extra-territorial organizations and bodies
- A 100 Agriculture, hunting and forestry
- B 200 Fishing
- C 300 Mining and quarrying
- D 400 Manufacturing
- E 500 Electricity, gas and water supply
- F 600 Construction
- G 700 Wholesale and retail trade; repair of motor Vehicles, motorcycles and personal and household goods
- H 800 Hotels and restaurants
- I 900 Transport, storage and communications
- J 1000 Financial intermediation
- K 1100 Real estate, renting and business activities
- L 1200 Public administration and defence; compulsory social security
- M 1300 Education
- N 1400 Health and social work
- O 1500 Other community, social and personal service activities
- P 1600 Private households with employed persons
- Q 1700 Extra-territorial organizations and bodies

7.2 國際標準職業分類 (ISCO-88)

1 LEGISLATORS, SENIOR OFFICIALS AND MANAGERS

11 Legislators and senior officials

111 Legislators

1110 Legislators

112 Senior government officials

1120 Senior government officials

113 Traditional chiefs and heads of villages

1130 Traditional chiefs and heads of villages

114 Senior officials of special-interest organisations

1141 Senior officials of political party organisations

1142 Senior officials of employers', workers' and other economic-interest organisations

1143 Senior officials of humanitarian and other special-interest organisations

12 Corporate managers

121 Directors and chief executives

1210 Directors and chief executives

122 Production and operations department managers

1221 Production and operations department managers in agriculture, hunting, forestry and fishing

1222 Production and operations department managers in manufacturing

1223 Production and operations department managers in construction

1224 Production and operations department managers in wholesale and retail trade

1225 Production and operations department managers in restaurants and hotels

1226 Production and operations department managers in transport, storage and communications

1227 Production and operations department managers in business services enterprises

1228 Production and operations department managers in personal care, cleaning and related services

1229 Production and operations department managers not elsewhere classified

123 Other department managers

1231 Finance and administration department managers

1232 Personnel and industrial relations department managers

1233 Sales and marketing department managers

1234 Advertising and public relations department managers

1235 Supply and distribution department managers

1236 Computing services department managers

1237 Research and development department managers

1239 Other department managers not elsewhere classified

1240 Office managers

13 General managers

131 General managers

1311 General managers in agriculture, hunting, forestry and fishing

1312 General managers in manufacturing

1313 General managers in construction

1314 General managers in wholesale and retail trade

1315 General managers of restaurants and hotels

1316 General managers in transport, storage and communications

1317 General managers in business services enterprises

1318 General managers in personal care, cleaning and related services

1319 General managers not elsewhere classified

2 PROFESSIONALS

21 Physical, mathematical and engineering science professionals

211 Physicists, chemists and related professionals

2111 Physicists and astronomers

2112 Meteorologists

2113 Chemists

2114 Geologists and geophysicists

212 Mathematicians, statisticians and related professionals

2121 Mathematicians and related professionals

2122 Statisticians

213 Computing professionals

- 2131 Computer systems designers, analysts and programmers
- 2132 Computer programmers
- 2139 Computing professionals not elsewhere classified
- 214 Architects, engineers and related professionals
 - 2141 Architects, town and traffic planners
 - 2142 Civil engineers
 - 2143 Electrical engineers
 - 2144 Electronics and telecommunications engineers
 - 2145 Mechanical engineers
 - 2146 Chemical engineers
 - 2147 Mining engineers, metallurgists and related professionals
 - 2148 Cartographers and surveyors
 - 2149 Architects, engineers and related professionals not elsewhere classified
- 22 Life science and health professionals
 - 221 Life science professionals
 - 2211 Biologists, botanists, zoologists and related professionals
 - 2212 Pharmacologists, pathologists and related professionals
 - 2213 Agronomists and related professionals
 - 222 Health professionals (except nursing)
 - 2221 Medical doctors
 - 2222 Dentists
 - 2223 Veterinarians
 - 2224 Pharmacists
 - 2229 Health professionals (except nursing) not elsewhere classified
 - 223 Nursing and midwifery professionals
 - 2230 Nursing and midwifery professionals
- 23 Teaching professionals
 - 2300 Teaching professionals (n.f.s.)
 - 231 College, university and higher education teaching professionals
 - 2310 College, university and higher education teaching professionals
 - 232 Secondary education teaching professionals
 - 2320 Secondary education teaching professionals
 - 2321 Secondary education teaching professionals in cram school
 - 233 Primary and pre-primary education teaching professionals
 - 2331 Primary education teaching professionals
 - 2332 Pre-primary education teaching professionals
 - 234 Special education teaching professionals
 - 2340 Special education teaching professionals
 - 235 Other teaching professionals
 - 2351 Education methods specialists
 - 2352 School inspectors
 - 2359 Other teaching professionals not elsewhere classified
- 24 Other professionals
 - 241 Business professionals
 - 2411 Accountants
 - 2412 Personnel and careers professionals
 - 2419 Business professionals not elsewhere classified
 - 242 Legal professionals
 - 2421 Lawyers
 - 2422 Judges
 - 2429 Legal professionals not elsewhere classified
 - 243 Archivists, librarians and related information professionals
 - 2431 Archivists and curators
 - 2432 Librarians and related information professionals
 - 244 Social science and related professionals
 - 2441 Economists
 - 2442 Sociologists, anthropologists and related professionals
 - 2443 Philosophers, historians and political scientists
 - 2444 Philologists, translators and interpreters
 - 2445 Psychologists

- 2446 Social work professionals
- 245 Writers and creative or performing artists
 - 2451 Authors, journalists and other writers
 - 2452 Sculptors, painters and related artists
 - 2453 Composers, musicians and singers
 - 2454 Choreographers and dancers
 - 2455 Film, stage and related actors and directors
- 246 Religious professionals
 - 2460 Religious professionals

3 TECHNICIANS AND ASSOCIATE PROFESSIONALS

31 Physical and engineering science associate professionals

- 311 Physical and engineering science technicians
 - 3111 Chemical and physical science technicians
 - 3112 Civil engineering technicians
 - 3113 Electrical engineering technicians
 - 3114 Electronics and telecommunications engineering technicians
 - 3115 Mechanical engineering technicians
 - 3116 Chemical engineering technicians
 - 3117 Mining and metallurgical technicians
 - 3118 Draughtspersons
 - 3119 Physical and engineering science technicians not elsewhere classified
- 312 Computer associate professionals
 - 3121 Computer assistants
 - 3122 Computer equipment operators
 - 3123 Industrial robot controllers
- 313 Optical and electronic equipment operators
 - 3131 Photographers and image and sound recording equipment operators
 - 3132 Broadcasting and telecommunications equipment operators
 - 3133 Medical equipment operators
 - 3139 Optical and electronic equipment operators not elsewhere classified
- 314 Ship and aircraft controllers and technicians
 - 3141 Ships' engineers
 - 3142 Ships' deck officers and pilots
 - 3143 Aircraft pilots and related associate professionals
 - 3144 Air traffic controllers
 - 3145 Air traffic safety technicians
- 315 Safety and quality inspectors
 - 3151 Building and fire inspectors
 - 3152 Safety, health and quality inspectors

32 Life science and health associate professionals

- 321 Life science technicians and related associate professional
 - 3211 Life science technicians
 - 3212 Agronomy and forestry technicians
 - 3213 Farming and forestry advisers
- 322 Modern health associate professionals (except nursing)
 - 3221 Medical assistants
 - 3222 Sanitarians
 - 3223 Dieticians and nutritionists
 - 3224 Optometrists and opticians
 - 3225 Dental assistants
 - 3226 Physiotherapists and related associate professionals
 - 3227 Veterinary assistants
 - 3228 Pharmaceutical assistants
 - 3229 Modern health associate professionals (except nursing) not elsewhere classified
- 323 Nursing and midwifery associate professionals
 - 3231 Nursing associate professionals
 - 3232 Midwifery associate professionals
- 324 Traditional medicine practitioners and faith healers
 - 3241 Traditional medicine practitioners
 - 3242 Faith healers

33 Teaching associate professionals

- 331 Primary education teaching associate professionals
 - 3310 Primary education teaching associate professionals
- 332 Pre-primary education teaching associate professionals
 - 3320 Pre-primary education teaching associate professionals
 - 333 Special education teaching associate professionals
 - 3330 Special education teaching associate professionals
- 334 Other teaching associate professionals
 - 3340 Other teaching associate professionals
 - 3341 Other teaching associate professionals of educational subjects

34 Other associate professionals

- 341 Finance and sales associate professionals
 - 3411 Securities and finance dealers and brokers
 - 3412 Insurance representatives
 - 3413 Estate agents
 - 3414 Travel consultants and organisers
 - 3415 Technical and commercial sales representatives
 - 3416 Buyers
 - 3417 Appraisers, valuers and auctioneers
 - 3418 Bank representatives *
 - 3419 Finance and sales associate professionals not elsewhere classified
- 342 Business services agents and trade brokers
 - 3421 Trade brokers
 - 3422 Clearing and forwarding agents
 - 3423 Employment agents and labour contractors
 - 3429 Business services agents and trade brokers not elsewhere classified
- 343 Administrative associate professionals
 - 3431 Administrative secretaries and related associate professionals
 - 3432 Legal and related business associate professionals
 - 3433 Bookkeepers
 - 3434 Statistical, mathematical and related associate professionals
 - 3439 Administrative associate professionals not elsewhere classified
- 344 Customs, tax and related government associate professionals
 - 3441 Customs and border inspectors
 - 3442 Government tax and excise officials
 - 3443 Government social benefits officials
 - 3444 Government licensing officials
 - 3449 Customs, tax and related government associate professionals not elsewhere classified
- 345 Police inspectors and detectives
 - 3450 Police inspectors and detectives
- 346 Social work associate professionals
 - 3460 Social work associate professionals
- 347 Artistic, entertainment and sports associate professionals
 - 3471 Decorators and commercial designers
 - 3472 Radio, television and other announcers
 - 3473 Street, night-club and related musicians, singers and dancers
 - 3474 Clowns, magicians, acrobats and related associate professionals
 - 3475 Athletes, sports persons and related associate professionals
- 348 Religious associate professionals
 - 3480 Religious associate professionals

4 CLERKS

41 Office clerks

- 411 Secretaries and keyboard-operating clerks
 - 4100 Office clerks (n.f.s.)
 - 4111 Stenographers and typists
 - 4112 Word-processor and related operators
 - 4113 Data entry operators
 - 4114 Calculating-machine operators
 - 4115 Secretaries

- 412 Numerical clerks
 - 4121 Accounting and book-keeping clerks
 - 4122 Statistical and finance clerks
- 413 Material-recording and transport clerks
 - 4131 Stock clerks
 - 4132 Production clerks
 - 4133 Transport clerks
- 414 Library, mail and related clerks
 - 4141 Library and filing clerks
 - 4142 Mail carriers and sorting clerks
 - 4143 Coding, proof-reading and related clerks
 - 4144 Scribes and related workers
- 419 Other office clerks
 - 4190 Other office clerks

42 Customer services clerks

- 421 Cashiers, tellers and related clerks
 - 4211 Cashiers and ticket clerks
 - 4212 Tellers and other counter clerks
 - 4213 Bookmakers and croupiers
 - 4214 Pawnbrokers and money-lenders
 - 4215 Debt-collectors and related workers
- 422 Client information clerks
 - 4221 Travel agency and related clerks
 - 4222 Receptionists and information clerks
 - 4223 Telephone switchboard operators

5 SERVICE WORKERS AND SHOP AND MARKET SALES WORKERS

51 Personal and protective services workers

- 511 Travel attendants and related workers
 - 5111 Travel attendants and travel stewards
 - 5112 Transport conductors
 - 5113 Travel guides
- 512 Housekeeping and restaurant services workers
 - 5121 Housekeepers and related workers
 - 5122 Cooks
 - 5123 Waiters, waitresses and bartenders
- 513 Personal care and related workers
 - 5131 Child-care workers
 - 5132 Institution-based personal care workers
 - 5133 Home-based personal care workers
 - 5139 Personal care and related workers not elsewhere classified
- 514 Other personal services workers
 - 5141 Hairdressers, barbers, beauticians and related workers
 - 5142 Companions and valets
 - 5143 Undertakers and embalmers
 - 5149 Other personal services workers not elsewhere classified
- 515 Astrologers, fortune-tellers and related workers
 - 5151 Astrologers and related workers
 - 5152 Fortune-tellers, palmists and related workers
- 516 Protective services workers
 - 5161 Fire-fighters
 - 5162 Police officers
 - 5163 Prison guards
 - 5164 Self-defense force personnel *
 - 5169 Protective services workers not elsewhere classified

52 Models, salespersons and demonstrators

- 521 Fashion and other models
 - 5210 Fashion and other models
- 522 Shop, stall and market salespersons and demonstrators
 - 5220 Shop, stall and market salespersons and demonstrators

- 523 Stall and market salespersons
 - 5230 Stall and market salespersons
- 524 Insurance Salespersons
 - 5240 Insurance Salespersons *

6 SKILLED AGRICULTURAL AND FISHERY WORKERS

61 Skilled agricultural and fishery workers

- 611 Market gardeners and crop growers
 - 6110 Market gardeners and crop growers (n.f.s.)
 - 6111 Field crop and vegetable growers
 - 6112 Tree and shrub crop growers
 - 6113 Gardeners, horticultural and nursery growers
 - 6114 Mixed-crop growers
- 612 Market-oriented animal producers and related workers
 - 6121 Dairy and livestock producers
 - 6122 Poultry producers
 - 6123 Apiarists and sericulturists
 - 6124 Mixed-animal producers
 - 6129 Market-oriented animal producers and related workers not elsewhere classified
- 613 Market-oriented crop and animal producers
 - 6130 Market-oriented crop and animal producers
- 614 Forestry and related workers
 - 6141 Forestry workers and loggers
 - 6142 Charcoal burners and related workers
- 615 Fishery workers, hunters and trappers
 - 6151 Aquatic life cultivation workers
 - 6152 Inland and coastal waters fishery workers
 - 6153 Deep-sea fishery workers
 - 6154 Hunters and trappers

62 Subsistence agricultural and fishery workers

- 621 Subsistence agricultural and fishery workers
 - 6210 Subsistence agricultural and fishery workers

7 CRAFT AND RELATED TRADES WORKERS

71 Extraction and building trades workers

- 711 Miners, shotfirers, stone cutters and carvers
 - 7111 Miners and quarry workers
 - 7112 Shotfirers and blasters
 - 7113 Stone splitters, cutters and carvers
- 712 Building frame and related trades workers
 - 7121 Builders, traditional materials
 - 7122 Bricklayers and stonemasons
 - 7123 Concrete placers, concrete finishers and related workers
 - 7124 Carpenters and joiners
 - 7129 Building frame and related trades workers not elsewhere classified
- 713 Building finishers and related trades workers
 - 7131 Roofers
 - 7132 Floor layers and tile setters
 - 7133 Plasterers
 - 7134 Insulation workers
 - 7135 Glaziers
 - 7136 Plumbers and pipe fitters
 - 7137 Building and related electricians
 - 7139 Building finisher (n.e.c.)
- 714 Painters, building structure cleaners and related trades workers
 - 7141 Painters and related workers
 - 7142 Varnishers and related painters
 - 7143 Building structure cleaners

72 Metal, machinery and related trades workers

- 721 Metal moulders, welders, sheet-metal workers, structural-metal preparers, and related trades workers
 - 7211 Metal moulders and coremakers

- 7435 Textile, leather and related pattern-makers and cutters
- 7436 Sewers, embroiderers and related workers
- 7437 Upholsterers and related workers
- 744 Pelt, leather and shoemaking trades workers
- 7441 Pelt dressers, tanners and fellmongers
- 7442 Shoe-makers and related workers
- 7499 Other craft & related trades workers (n.e.c.)
- 7510 Non-farm Manual Foremen and Supervisors (n.f.s.)

8 PLANT AND MACHINE OPERATORS AND ASSEMBLERS

- 8000 Plant and machine operators and assemblers (n.f.s.)

81 Stationary plant and related operators

- 811 Mining and mineral-processing-plant operators
 - 8111 Mining plant operators
 - 8112 Mineral-ore and stone-processing-plant operators
 - 8113 Well drillers and borers and related workers
- 812 Metal-processing plant operators
 - 8121 Ore and metal furnace operators
 - 8122 Metal melters, casters and rolling-mill operators
 - 8123 Metal heat-treating-plant operators
 - 8124 Metal drawers and extruders
- 813 Glass, ceramics and related plant operators
 - 8131 Glass and ceramics kiln and related machine operators
 - 8139 Glass, ceramics and related plant operators not elsewhere classified
- 814 Wood-processing- and papermaking-plant operators
 - 8141 Wood-processing-plant operators
 - 8142 Paper-pulp plant operators
 - 8143 Papermaking-plant operators
- 815 Chemical-processing-plant operators
 - 8151 Crushing-, grinding- and chemical-mixing-machinery operators
 - 8152 Chemical-heat-treating-plant operators
 - 8153 Chemical-filtering- and separating-equipment operators
 - 8154 Chemical-still and reactor operators (except petroleum and natural gas)
 - 8155 Petroleum- and natural-gas-refining-plant operators
 - 8159 Chemical-processing-plant operators not elsewhere classified
- 816 Power-production and related plant operators
 - 8161 Power-production plant operators
 - 8162 Steam-engine and boiler operators
 - 8163 Incinerator, water-treatment and related plant operators
- 817 Automated-assembly-line and industrial-robot operators
 - 8171 Automated-assembly-line operators
 - 8172 Industrial-robot operators

82 Machine operators and assemblers

- 821 Metal- and mineral-products machine operators
 - 8211 Machine-tool operators
 - 8212 Cement and other mineral products machine operators
- 822 Chemical-products machine operators
 - 8221 Pharmaceutical-and toiletry-products machine operators
 - 8222 Ammunition- and explosive-products machine operators
 - 8223 Metal finishing-, plating- and coating-machine operators
 - 8224 Photographic-products machine operators
 - 8229 Chemical-products machine operators not elsewhere classified
- 823 Rubber- and plastic-products machine operators
 - 8231 Rubber-products machine operators
 - 8232 Plastic-products machine operators
- 824 Wood-products machine operators
 - 8240 Wood-products machine operators
- 825 Printing-, binding- and paper-products machine operators
 - 8251 Printing-machine operators
 - 8252 Book-binding-machine operators
 - 8253 Paper-products machine operators

- 826 Textile-, fur- and leather-products machine operators
 - 8261 Fibre-preparing-, spinning- and winding-machine operators
 - 8262 Weaving- and knitting-machine operators
 - 8263 Sewing-machine operators
 - 8264 Bleaching-, dyeing- and cleaning-machine operators
 - 8265 Fur- and leather-preparing-machine operators
 - 8266 Shoemaking- and related machine operators
 - 8269 Textile-, fur- and leather-products machine operators not elsewhere classified
- 827 Food and related products machine operators
 - 8270 Food and related products machine operators (n.f.s.)
 - 8271 Meat- and fish-processing-machine operators
 - 8272 Dairy-products machine operators
 - 8273 Grain- and spice-milling-machine operators
 - 8274 Baked-goods, cereal- and chocolate-products machine operators
 - 8275 Fruit-, vegetable- and nut-processing-machine operators
 - 8276 Sugar production machine operators
 - 8277 Tea-, coffee- and cocoa-processing-machine operators
 - 8278 Brewers, wine and other beverage machine operators
 - 8279 Tobacco production machine operators
- 828 Assemblers
 - 8281 Mechanical-machinery assemblers
 - 8282 Electrical-equipment assemblers
 - 8283 Electronic-equipment assemblers
 - 8284 Metal-, rubber- and plastic-products assemblers
 - 8285 Wood and related products assemblers
 - 8286 Paperboard, textile and related products assemblers
 - 8287 Composite products assemblers
- 829 Other machine operators and assemblers
 - 8290 Other machine operators and assemblers
- 83 Drivers and mobile-plant operators**
 - 831 Locomotive-engine drivers and related workers
 - 8311 Locomotive-engine drivers
 - 8312 Railway brakemen, signallers and shunters
 - 832 Motor-vehicle drivers
 - 8321 Motor-cycle drivers
 - 8322 Car, taxi and van drivers
 - 8323 Bus and tram drivers
 - 8324 Heavy truck and lorry drivers
 - 833 Agricultural and other mobile plant operators
 - 8330 Agricultural and other mobile plant operators (n.f.s.)
 - 8331 Motorised farm and forestry plant operators
 - 8332 Earth-moving and related plant operators
 - 8333 Crane, hoist and related plant operators
 - 8334 Lifting-truck operators
 - 834 Ships' deck crews and related workers
 - 8340 Ships' deck crews and related workers
- 9 ELEMENTARY OCCUPATIONS**
 - 9000 Elementary occupations (n.f.s.)
- 91 Sales and services elementary occupations**
 - 911 Street vendors and related workers
 - 9111 Street food vendors
 - 9112 Street vendors, non-food products
 - 9113 Door-to-door salespersons
 - 9114 Salesperson Via Telecommunication
 - 912 Shoe cleaning and other street services elementary occupations
 - 9120 Shoe cleaning and other street services elementary occupations
 - 913 Domestic and related helpers, cleaners and launderers
 - 9131 Domestic helpers and cleaners
 - 9132 Helpers and cleaners in offices, hotels and other establishments
 - 9133 Hand-launderers and pressers

- 914 Building caretakers, window and related cleaners
 - 9141 Building caretakers
 - 9142 Vehicle, window and related cleaners
- 915 Messengers, porters, doorkeepers and related workers
 - 9151 Messengers, package and luggage porters and deliverers
 - 9152 Doorkeepers, watchpersons and related workers
 - 9153 Vending-machine money collectors, meter readers and related workers
- 916 Garbage collectors and related labourers
 - 9161 Garbage collectors
 - 9162 Sweepers and related labourers
- 92 Agricultural, fishery and related labourers**
 - 921 Agricultural, fishery and related labourers
 - 9211 Farm-hands and labourers
 - 9212 Forestry labourers
 - 9213 Fishery, hunting and trapping labourers
- 93 Labourers in mining, construction, manufacturing and transport**
 - 931 Mining and construction labourers
 - 9311 Mining and quarrying labourers
 - 9312 Construction and maintenance labourers: roads, dams and similar constructions
 - 9313 Building construction labourers
 - 932 Manufacturing labourers
 - 9321 Assembling labourers
 - 9322 Hand packers and other manufacturing labourers
 - 933 Transport labourers and freight handlers
 - 9331 Hand or pedal vehicle drivers
 - 9332 Drivers of animal-drawn vehicles and machinery
 - 9333 Freight handlers
- 0 ARMED FORCES**
 - 01 Armed forces**
 - 010 Armed forces
 - 0100 Armed forces

(注) *印は日本の事情に合わせるために作成した新コード。その他の新コードについてはコーディング・ガイドを参照されたい。

8. ISIC, ISCO のコーディング・ガイド

2005 年 SSM 日本調査では、現職、職歴、父・母職、配偶者職のすべてについて、SSM 産業、職業分類に加えて、国際標準産業分類（ISIC）、国際標準職業分類（ISCO）をコードした。

8.1 国際標準産業分類（ISIC）

1. ISIC の構造

(1) ISIC とは

ISIC（International Standard Industrial Classification of all economic activities）とは、United Nations において作成された産業分類で、すべての生産的な経済活動（economic productive activity）を分類することを目的としている。1948 年にはじめて作成された後、何度か改訂され、現在の第 3 改訂版（ISIC Rev3）は、各国の産業分類のモデルとなっている。

(2) ISIC コードの概説

ISIC コードは、分類の詳細度が 4 段階に分かれている。一番大きな分類はアルファベットで構成されている（大分類）。次に細かいレベルの分類は 2 桁の数字で分類される（2 桁分類）。更に細かいレベルの分類は 3 桁目まで表示され（3 桁分類）、一番細かいレベルの分類は 4 桁数字によって表示される（4 桁分類）。表 1 は、ISIC コードの大分類表である。なお、アルファベットは以下のような数値に変換した。

表 1 ISIC コード

大分類	コード	分類名(英語)	分類名(日本語)
A	100	Agriculture, hunting and forestry	農林狩猟業
B	200	Fishing	漁業
C	300	Mining and quarrying	鉱業と採石業
D	400	Manufacturing	製造業
E	500	Electricity, gas and water supply	電気、ガスと給水業
F	600	Construction	建設業
G	700	Wholesale and retail trade; repair of motor Vehicles, motorcycles and personal and household goods	卸売業、小売業、自動車・バイク・家電修理業
H	800	Hotels and restaurants	ホテル・レストラン業
I	900	Transport, storage and communications	運輸・倉庫・通信業
J	1000	Financial intermediation	金融仲介業
K	1100	Real estate, renting and business activities	不動産業・賃貸業
L	1200	Public administration and defense; compulsory social security	公務及び国防(社会保障)
M	1300	Education	教育
N	1400	Health and social work	保健と社会福祉業
O	1500	Other community, social and personal service activities	他のコミュニティ、社会的および個人サービス業
P	1600	Private households with employed persons	家事サービス業
Q	1700	Extra-territorial organizations and bodies	治外法権組織

2. コーディング・ルール

(1) 一般的なルール

今回の産業コーディング作業では、主に2桁分類を用いる。しかしながら、ISICの2桁分類と『SSM産業分類』は、必ずしも対応しているわけではない。例えば、ISIC2桁分類では「55 Hotels and restaurants」には『SSM産業分類』の「93 飲食店」とホテル業（SSM産業分類では「17 その他サービス」）が含まれる。そこで、ISICと『SSM産業分類』を対応させるために、ISICコードは2桁ではなく、3桁あるいは4桁コードを用いる場合もある。例えば、先の例であれば、「55 Hotels and restaurants」を3桁分類にまで細分化し、「飲食店」ならば、ISIC3桁分類の「552 Restaurants, bars and canteens」を用い、「ホテル業」ならば、「551 Hotels; camping sites and other provision of short-stay accommodation」を用いる。

(2) 特記事項

・複数の産業が記載されている場合

1. 主要な産業がどちらかわかる場合には、主要な産業の方をコード。
2. 「製造業」が含まれれば、製造を優先。

例えば、「プラスチック製品の販売製造会社」ならば、「25 Manufacture of rubber and plastics products」。

3. （上記2つのルールでも決まらない場合）最初に記述されているものを優先する

3. 新設コードとそのコード番号について

・塾産業・・・「804 Tutorial School」

「塾産業」に関しては、2005年SSM韓国調査で使用している産業分類（KSIC）との対応を考慮し、新コードを作成した。

・学校給食・・・「805 School lunch」

「学校給食」に関しては、『SSM産業分類』とKSIC、両方との対応が可能となるように新コードを作成した。

4. 2桁分類よりも細密な分類を用いたケースについて

今回の産業コーディングでは、主に2桁分類を用いているが、ISICの2桁分類と『SSM産業分類』が対応しないケースについては、より細かな分類を用いた。詳細は以下の通りである。

(1) 「22 Publishing, printing and reproduction of recorded media」

製造業と出版業が混在しているため3桁分類を使用した。

- ・221 Publishing
- ・222 Printing and service activities related to printing
- ・223 Reproduction of recorded media

(2) 「35 Manufacture of other transport equipment」

35の中に製造と修理が入るため、その他サービスにあたる修理業を分けた。

- ・35 Manufacture of other transport equipment・・・その他輸送機器製造業
- ・351 Building and repairing of ships・・・船舶の製造および修理

(3) 「50 Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel」

販売と修理が混在しているため3桁分類を用いた。

- ・501 Sale of motor vehicles・・・販売
- ・502 Maintenance and repair of motor vehicles・・・整備、修理
- ・503 Sale of motor vehicle parts and accessories・・・販売
- ・504 Sale, maintenance and repair of motorcycle and related parts and accessories・・・販売、整備
- ・505 Retail sale of automotive fuel・・・販売

(4) 「52 Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods」

52の中に販売と修理が入るため、その他サービスにあたる修理業を分けた。

- ・ 52 Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods・・・販売
- ・ 526 Repair of personal and household goods・・・修理
- (5) 「55 Hotels and restaurants」
ホテル・旅館業と飲食業が混在しているため3桁分類を用いた。
 - ・ 551 Hotels; camping sites and other provision of short-stay accommodation・・・ホテル、旅館
 - ・ 552 Restaurants, bars and canteens・・・飲食
- (6) 「64 Post and telecommunications」
64の郵便事業の中に宅配業も入るため、それらを4桁分類により分けた。
 - ・ 64 Post and telecommunications・・・郵便事業
 - ・ 6412 Courier activities other than national post activities・・・郵便事業を除く宅配業
- (7) 「72 Computer and related activities」
72の中に情報処理とコンピュータメンテナンスが入るため、通信とサービスを分けた。
 - ・ 72 Computer and related activities・・・コンピュータ関連、データ通信事業
 - ・ 725 Maintenance and repair of office, accounting and computing machinery・・・修理業
- (8) 「74 Other business activities」
74の中にビジネスサービスと法律、会計業務、広告業務が入っているため、それらを分類した。
 - ・ 74 Other business activities・・・ビジネスサービス
 - ・ 7411 Legal activities・・・法律業務
 - ・ 7412 Accounting, book-keeping and auditing activities; tax consultancy・・・会計業務
 - ・ 743 Advertising
- (9) 「80 Education」
教育産業から塾、学校給食を分類した（新設コード）。
 - ・ 80 Education・・・教育
 - ・ 804 Tutorial school・・・塾、受験産業
 - ・ 805 School lunch・・・学校給食
- (10) 「85 Health and social work」
医療、保健、福祉業務から、はり・灸・あんま、獣医を分類した。
 - ・ 85 Health and social work・・・医療、福祉、保健
 - ・ 8519 Other human health activities・・・はり・灸・あんま
 - ・ 8520 Veterinary activities・・・獣医
- (11) 「92 Recreational, cultural and sporting activities」
92は娯楽産業、マスメディア、図書館・博物館が混在しているため、3桁分類を用いた。
 - ・ 921 Motion picture, radio, television, and other entertainment activities・・・映画、ラジオ、テレビ、芸能
 - ・ 922 News agency activities・・・新聞、報道
 - ・ 923 Library, archives, museums and other cultural activities・・・図書館、博物館、動物園
 - ・ 924 Sporting and other recreational activities・・・スポーツ、娯楽、趣味の学校

8.2 国際標準職業分類（ISCO-88）

1. ISCO-88 の構造

(1) ISCO-88 とは

ISCO（International Standard Classification of Occupation）とは国際労働機構（ILO）が作成した職業分類で、各国の職業分類のモデルとなっている

(2) ISCO-88 の概念的フレームワーク

1 個人の果たす「job」（＝a set of tasks and duties executed）を分類する。

2 スキルレベル（＝教育・職業資格）によって大分類が異なる。

ある職業が果たすタスクや義務を行うのに必要とされる「スキル」を考慮する。

（同じ職業に従事する労働者間のスキルの違いなどは考慮しない）

3 マニュアル職の場合はスキルレベルのほかに、「手作業」か「機械操作」かで区別する。

同じスキルレベル2でも、大分類7は「手作業・職人的なスキル」であり、大分類8は「機械操作・流れ作業的なスキル」となる。

表1 ISCO-88 の大分類とそのスキルレベル、分類数

大分類	大分類名	スキルレベル	亜大分類	中分類	小分類
1	Legislators, senior officials and managers 管理職	－	3	8	33
2	Professionals 専門職	4	4	18	55
3	Technicians and associate professionals 技術者・準専門職	3	4	21	73
4	Clerks 事務職	2	2	7	23
5	Service workers and shop and market sales workers サービス職・販売職	2	2	9	23
6	Skilled agricultural and fishery workers 農林漁業従事者	2	2	6	17
7	Craft and related trade workers マニュアル(手作業・職人系)	2	4	16	70
8	Plant and machine operators and assemblers マニュアル(機械操作・組立)	2	3	20	70
9	Elementary occupations マニュアル(労務・単純作業)	1	3	10	25
0	Armed force 軍人	－	1	1	1

注1) 大分類1（管理職）と大分類0（軍人）はスキルレベルにとらわれず分類。

注2) 大分類0（軍人）は日本では旧軍の軍人にのみに使う。

表2 スキルレベルと学歴の対応

スキルレベル	学歴対応
4	大学(4年生)卒業以上
3	短大・専修学校卒業
2	中学・高校卒業以上
1	小学校卒業以上

(3) ISCO コードの概説

ISCO コードは4桁からなっており、

1桁目が大分類、2桁で亜大分類、3桁目までで中分類、4桁で小分類を示す。

例えば、大工（carpenters）は、

大分類 7 （Craft and related trade workers）

- 亜人分類 71 (Extraction and building trade workers)
- 中分類 712 (Building frame and related trade workers)
- 小分類 7124 (Carpenters and joiner)

とたどると、具体的なコードをあてはめることができる。

また 2 桁目以降の数字の意味として、「0」と「9」がそれぞれ

0=n.f.s. (=not further specified) = 「これ以上詳しい分類ができない」

9=n.e.c. (=not elsewhere classified) = 「その他に含まれない」 を意味する。

例：「2300」(Teaching professionals, n.f.s.) などがある。

仕事の内容として「教員」とのみ記載され、小学校の教員「2331」なのか、中学や高校の教員「2320」なのかを区別する情報がない場合、中分類の 23 (Teaching Professionals) に n.f.s. として 00 を加えた「2300」というコードを作成することもできる。
これにより、新コード作成を体系的に行うことができる。

2. 一般的なコーディング・ルール

(1) ISCO-88 コードブックのルールより

- ・ 技術者 (technicians) → 基本的に 3000 番台にコードする
- ・ 品質検査の仕事について
(例：「プラスチック部品の検査」、「食品の安全基準の審査・検査」など)
品質規格や製品仕様が遵守されているかを確認する仕事
→ 「3152」(Safety, health and quality inspectors) にコードする
テスターやチェッカーとして機械的に点検しているだけ→各製造工程にコード
- ・ 指導 (coaching) の仕事は、指導する業務の内容に従ってコード
- ・ 研究や開発に関わる仕事は、その専門に応じた専門職(2000 番台)にする
ただし教える仕事もしている場合は、その教育レベルに合わせた教師としてコード
例：「大学で研究、授業の講義」ならば
→ 「2310」(College, university and higher education teaching professionals)

(2) コーディング用のコード番号

- ・ 無回答は (99999) とコードし、分類不能は (0) とコードする
職業に就いてはいるが、何も具体的な回答がない場合は「無回答」。
一方、記述があるが詳細がわからないことなどから分類できない場合「分類不能」

(3) 「管理職」の取り扱いについて

以下の①～③の順に優先してコードする。

(①で対応不可能なら②を、②でも対応できない場合に③を、という順で用いる)

- ① 「仕事の内容」が「管理」や「経営」など管理者としての記載しかない場合のみ使用する。
(役職が「部長」など管理職、あるいは従業上の地位が「経営者」などであっても「仕事の内容」に管理的仕事以外の内容が含まれていれば、その内容を優先してコードする。)
- ② 記述内容に「管理」と書いてあっても、資材や製品など「人」以外を「管理」している場合は、「事務職」と考えて管理職としては扱わない。
- ③ 「管理職」としてコードする場合、「建設系」とそれ以外で取り扱いが異なる。

A. 産業あるいは仕事の内容が「土木・建築・建設の仕事」や「土建業」などの場合

- (a) 役職が「監督、職長、班長、組長」または「係長、係長相当職」の場合
 - ・ 規模 30 人未満 かつ 学歴が大卒未満

「7510」(Non-farm Manual Foremen and Supervisors n.f.s.)

- ・規模 30 人以上 かつ 学歴が大卒以上

「3112」(Civil engineering technicians)

(b) 役職が「課長、課長相当職」または「部長、部長相当職」の場合

- ・規模 30 人未満 「7510」(Non-farm Manual Foremen and Supervisors n.f.s.)
- ・規模 30 人以上 「1223」(Production and operations department managers in construction)
- ・規模が不明 「1240」(Office managers) という新コードを使う

(c) 役職が「経営者・役員」の場合

- ・規模 30 人未満 「1313」(General managers in construction) にコード
- ・規模 30 人以上 「1210」(Directors and chief executives)
- ・規模が不明 「1313」(General managers in construction)

B. 産業あるいは仕事内容が「建築系」以外の場合

(a) 役職が「課長、課長相当職」または「部長、部長相当職」の場合、

- ・規模 30 人未満 「13XX」(Small Enterprise General Manager in XX) を使う
- ・規模 30 人以上 「12XX」(Large Enterprise Corporate Manager in XX) を使う
- ・規模が不明 「1240」(Office managers) という新コードを使う

(b) 役職が「経営者・役員」の場合、

- ・規模 30 人未満 「13XX」(Small Enterprise General Manager in XX) を使う
- ・規模 30 人以上 「1210」(Directors and chief executives)
- ・規模が不明 「13XX」(Small Enterprise General Manager in XX)

(4) 複数の仕事内容が記載されている場合

(「おべんとう作りと、その販売」などのように仕事内容が複数の職種にまたがる場合)
以下の①～④のルールを、順に優先してコードする。

(① で対応不可能なら②を、②でも対応できない場合に③を、という順で用いる)

① 主要なタスクでコードする

例：「現場で大工として工事全般とそれに関わる事務作業」

→「事務作業」は「工事」に付帯する作業と考え、この場合は「大工」にコード

② スキルレベルの高い方にコードする

例：「給食作り、栄養士」ならば「栄養士」をコードする

給食作り=5122 (Cooks)

栄養士 =3223 (Dieticians and nutritionists) 栄養士の方がスキルレベルが高い

③ 製造を（それに伴っての販売などより）優先する

例：「おべんとう作りと、その販売」ならば「おべんとう作り」にコード

販売 =5220 (Shop, stall and market salespersons and demonstrators)

弁当作り=7419 (Food processing & related trade workers n.e.c.)

→スキルレベルはともに「2」で等しいので、製造を優先して 7419 にコード

④ 上の 3 つで決められない場合、最初を書いてあるものを優先する

例：「雑用、経理事務、タイピング」ならば、「経理事務」にコード

雑用 =9000 (Elementary occupations n.f.s.)

経理事務=4121 (Accounting and book-keeping clerks)

タイピング=4111 (Stenographers and typists)

→スキルレベルから雑用「9000」ではなく、経理事務「4121」かタイピング「4111」

次に④の基準を利用して、先に出てある方を優先し「4121」にコードする

3. 新設コードとそのコード番号について

(1) Ganzeboom & Treiman (1996) 掲載の新コード一覧

- 1200～ [Large Enterprise]～ (元のコードの名称に追加)
 1300～ [Small Enterprise]～ (元のコードの名称に追加)
 1240 Office managers オフィスの管理職
 7510 Non-farm Manual Foremen and Supervisors n.f.s. マニュアル職の現場監督
 (これ以上分類できない)

(2) KSCO などから採用した新設コード

- 2321: Secondary education teaching professionals in cram school 予備校の教師
 3341: Other teaching associate professionals of educational subjects 学校教科教育の準専門職
 ・・・・「塾」(特に学習塾など)の先生や家庭教師などに使用
 7139: Building finisher (n.e.c.) 建築仕上げ工 (その他に分類できない)
 9114: Salesperson Via Telecommunication 通信機器を通じた販売員

(3) n.f.s. (これ以上分類できない) による新設コード

- 2300: Teaching professionals(n.f.s.) 教員 (これ以上分類できない)
 4100: Office clerks(n.f.s.) 事務所の事務 (これ以上分類できない)
 6110: Market gardeners and crop growers 市場向け園芸家、作物栽培員 (これ以上分類できない)
 ・・・・「農業」とのみ記載され、耕作物の内容が分からない場合
 7410: Food processing & related trade workers(n.f.s.) 食品加工および関連職従事者
 (これ以上分類できない)
 ・・・・「食品製造」との記載のみで材料が不明の場合 (手作業)
 8270: Food and related products machine operators(n.f.s.) 食物および関連製品製造機械操作員
 (これ以上分類できない)
 ・・・・「食品製造」との記載のみで材料が不明の場合に (機械作業)
 8330: Agricultural and other mobile plant operators(n.f.s.) 農業用その他の移動プラント操作員
 (これ以上分類できない)
 ・・・・「重機運転」という記述のみで重機の種類が分からない場合
 9000: Elementary occupations(n.f.s.) 単純作業 (これ以上分類できない)
 ・・・・「雑役」「労務」など内容が不明な労務作業に使用

(4) n.e.c. (その他に含まれない) による新設コード

- 7419: Food processing & related trade workers(n.e.c.) その他の食品加工および関連職従事者
 ・・・・「弁当製造」など複数の素材を用いる食品製造に使用
 7499: Other craft & related trades workers(n.e.c.) その他の職人および関連職従事者
 ・・・・素材がわからないモノの製造に使用 (ex.義足製造)

(5) 日本の事情に合わせるため、コード作業から必要上作成したもの

- 3418: Bank representatives 銀行の外交員
 5164: Self-defense force personnel 自衛隊員
 5240: Insurance Salespersons 保険の外交員

[参考文献]

- International Labour Office, 1990, *International Standard Classification of Occupations: ISCO-88*, International Labour Office.
 Ganzeboom H.B.G., and D.J. Treiman, 1996, "Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations," *Social Science Research*, 25(3): 201-239.

Automatic Occupation Coding with Combination of Machine Learning and Hand-Crafted Rules

KAZUKO TAKAHASHI¹,
HIROYA TAKAMURA², and MANABU OKUMURA²

¹ KEIAI UNIVERSITY, FACULTY OF INTERNATIONAL STUDIES

² TOKYO INSTITUTE OF TECHNOLOGY,
PRECISION AND INTELLIGENCE LABORATORY

PACIFIC-ASIA KNOWLEDGE DISCOVERY AND DATA MINING (PAKDD-05)

Abstract

- We apply a machine learning method to an occupation coding in social surveys.
 - Support Vector Machines (SVMs)
 - Their combinations with hand-crafted rules
- We show that SVMs outperform the rule-based method.
- We show that the combination of two methods yields better accuracy.

Table of Contents

1. The Occupation Coding
2. Rule-Based Method
3. Machine Learning Method
 - 3.1 Application of SVMs
 - 3.2 The Combinations of SVMs and Rule-Based Method
4. Experiments
5. Conclusion

Table of Contents

1. The Occupation Coding
2. Rule-Based Method
3. Machine Learning Method
 - 3.1 Application of SVMs
 - 3.2 The Combinations of SVMs and Rule-Based Method
4. Experiments
5. Conclusion

The Occupation Coding

Occupation Data  Occupation Code

- job task (open-ended) nearly 200 categories
- industry (open-ended)
- employment
- job title
- firm size

Example

Occupation Data

- job task to arrange the delivery vehicles
- industry load and unload of luggage
- employment 2 : Regular employee
- job title 1 : No managerial post
- firm size 8 : From 500 to 999



Occupation Code 563 (a transportation clerk)

Problem

- The task is complicated.
- The number of responses is large.
- ↓
- Coder's labor is huge.
- ↓
- The results are sometimes inconsistent.

Table of Contents

1. The Occupation Coding
2. Rule-Based Method
3. Machine Learning Method
 - 3.1 Application of SVMs
 - 3.2 The Combinations of SVMs and Rule-Based Method
4. Experiments
5. Conclusion

Process

1. Extract a triplet **<verb, case, noun>** from “job task” and “industry”
2. Generalize the verb to a verb class
3. Search for **a rule** that matches the generalized triplet

Performance of the rule-based method

	(%)		
	JGSS-2000	JGSS-2001	JGSS-2002
Total number of samples	6,848	6,448	6,770
Total accuracy	67.3	65.8	66.1
Accuracy for the label-assigned samples	80.9	79.7	79.8
Coders' total accuracy	68.8 - 80.0%		

Accuracy

Accuracy in this presentation :

$$\frac{\text{the number of correctly-classified samples}}{\text{the number of all samples}}$$

Performance of the rule-based method

	(%)		
	JGSS-2000	JGSS-2001	JGSS-2002
Total number of samples	6,848	6,448	6,770
Total accuracy	67.3	65.8	66.1
Accuracy for the label-assigned samples	80.9	79.7	79.8
Coders' total accuracy	68.8 - 80.0%		

Problems

- We have to make a constant effort to maintain both the rule-set and the thesaurus
- The system cannot deal with the responses that cannot be transformed into the form of case frames.
(We call them *undetermined samples*)



The categorization performance is not satisfactory.

Table of Contents

1. The Occupation Coding
2. Rule-Based Method
3. Machine Learning Method
 - 3.1 Application of SVMs
 - 3.2 The Combinations of SVMs and Rule-Based Method
4. Experiments
5. Conclusion

Machine Learning Method (1/2)

Support Vector Machines (SVMs)

1. Create basic features from responses
 - words in responses to “job task”
 - words in responses to “industry”
 - responses to “employment status” and “job title”
2. Train SVMs
3. Determine occupation codes of test samples

Machine Learning Method (2/2)

one-versus-rest method to extend SVMs to multiclass-classifier

Table of Contents

1. The Occupation Coding
2. Rule-Based Method
3. Machine Learning Method
 - 3.1 Application of SVMs
 - 3.2 The Combinations of SVMs and Rule-Based Method
4. Experiments
5. Conclusion

The Combinations of SVMs and Rule-Based Method (1/2)

1. Add new features for SVMs
 - **add-code**
occupation codes provided by the rule-based method
 - **add-rule**
rules used in the rule-based method
 - **add-code-rule**
occupation codes + **rules**

The Combinations of SVMs and Rule-Based Method (2/2)

2. Use SVMs only when the rule-based method cannot determine a unique occupation code
 - **seq** (sequential applied)

Example (1/2)

- the rule-based method
- **add-rule**
 $\text{feature}_1 \dots \text{feature}_n$ **feature_{n+1}**
 (basic features)
 - **add-code**
 $\text{feature}_1 \dots \text{feature}_n$ **feature_{n+1}** **occupation**
 (basic features) code j
- rule i**

Example (2/2)

- **seq**
- | | the rule-based method | SVMs | result |
|-------|-----------------------|-------|--------|
| Case1 | 601 | 501 → | 601 |
| Case2 | undetermined (999) | 501 → | 501 |
| Case3 | 501, 502 | 501 → | 501 |

Table of Contents

1. The Occupation Coding
2. Rule-Based Method
3. Machine Learning Method
 - 3.1 Application of SVMs
 - 3.2 The Combinations of SVMs and Rule-Based Method
4. Experiments
5. Conclusion

The purpose of Experiments

- **Experiment 1**
 - Compare SVMs with the rule-based method
 - Investigate effective combinations of SVMs and the rule-based methods
- **Experiment 2**
 - Investigate the relationship between the size of a training data set and categorization accuracy

Data set

- JGSS-2000 (6,848 samples)
 - JGSS-2001 (6,448 samples)
 - JGSS-2002 (6,770 samples)
- | | |
|-------|----------------|
| total | 20,066 samples |
|-------|----------------|

Basic features

- words in responses to “job task”
- words in responses to “industry”
- responses to “employment status” and “job title”

The results of preliminary experiments

10-fold cross validation

- “firm size” + basic features
- 2-gram/3-gram + basic features
- Kana-basic features
- 2-gram/3-gram + kana-basic features
- Features selection by Information Gain
- Changing the dimension of the linear kernel

No significant improvement

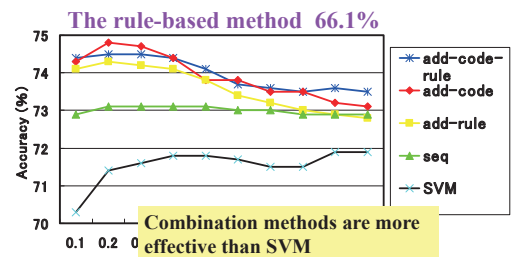
Experiment 1

• Data set

	Data set	Number of samples
Training data	JGSS-2000 JGSS-2001	13,296
Test data	JGSS-2002	6,770

- Soft margin parameter C : 0.1 - 1.0

Accuracy of each method with different



Accuracy of SVM is higher than that of the rule-based method

Tuning the value of C

Temporary training data : JGSS-2000

Temporary test data : JGSS-2001

	SVM	add-code	add-rule	add-code-rule
Predicted best C Accuracy on JGSS-2002	0.6 71.7	0.4 74.4	0.3 74.2	0.2 74.5
Actual best C Accuracy on JGSS-2002	1.0 71.9	0.2 74.8	0.2 74.3	0.2 74.5

Experiment 2

The coded samples of the previous surveys **available** (Case 1) or **not available** (Case 2)

• Training data set

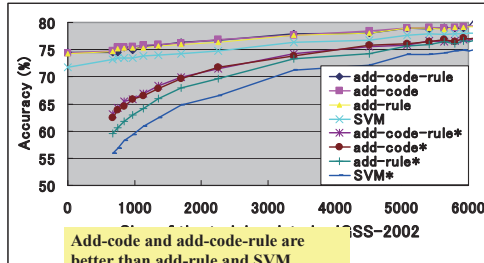
- Case 1
JGSS-2000+JGSS-2001 + a part of JGSS-2002
- Case 2
a part of JGSS-2002

• Test data set

the rest of JGSS-2002

Re size of training data

Accuracy in Case 2 with the half of the newly-added training data approximately equals to accuracy in Case 1 without any newly-added training samples



Accuracy increases as a size of a training data set becomes larger

Add-code and add-code-rule are better than add-rule and SVM

To sum up

Experiment 1

- The ranking of the automatic methods is **add-code-rule** nearly equal to **add-code**(75%) > **add-rule**(74%) > **seq**(73%) > **SVM**(72%) >> the rule-based method(66%)

Experiment 2

- The larger the training dataset, the higher the accuracies, for each method.
- The ranking is stable.
- Without coded samples, conducting by the half size of all the samples is effective.

Table of Contents

1. The Occupation Coding
2. Rule-Based Method
3. Machine Learning Method
 - 3.1 Application of SVMs
 - 3.2 The Combinations of SVMs and Rule-Based Method
4. Experiments
5. Conclusion

Conclusions

- We have applied **SVMs** to the occupation coding and shown that **SVMs** are superior to the rule-based method in terms of categorization accuracy also when a document is very short.
- We have also applied **the combinations of SVMs and the rule-based method** to the occupation coding and shown that **each of the combination methods** is superior to SVMs.
- We have shown that **a feedback** is effective.

Future Work

1. We would like to find a method for measuring confidence for each output of these automatic methods.
2. We will also adopt active learning in the feedback process.

Thank you !

Estimation of Class Membership Probabilities in the Document Classification

KAZUKO TAKAHASHI¹,
HIROYA TAKAMURA², and MANABU OKUMURA²

¹ KEIAI UNIVERSITY, FACULTY OF INTERNATIONAL STUDIES

² TOKYO INSTITUTE OF TECHNOLOGY,
PRECISION AND INTELLIGENCE LABORATORY

PACIFIC-ASIA KNOWLEDGE DISCOVERY AND DATA MINING (PAKDD-07)

1

Table of Contents

1. Motivation
2. Proposed Method
 - a. A Method Using an Accuracy Table
 - b. A Method Applying a Logistic Regression
3. Experiments
4. Conclusions

2

Table of Contents

1. **Motivation**
2. **Proposed Method**
 - a. A Method Using an Accuracy Table
 - b. A Method Applying a Logistic Regression
3. Experiments
4. Conclusions

3

Motivation

Estimating the probability with which the sample belongs to the predicted class (**class membership probability**) is useful in many applications such as document classification.

◆ Human decision making

e.g. the NANACO system

displays outputs from the automatic system as candidates of occupational codes to help human annotators (coders) for the occupation coding in social surveys

with **class membership probabilities**

4

The Occupation Coding

Occupation Data → Occupational Code

- job task (open-ended)
 - industry (open-ended)
 - employment
 - job title
 - firm size
- one of nearly 200 categories

5

A Picture by the NANACO System

Class Membership Probability

Occupation Data **Candidates of Occupational Codes**

本人職業 to arrange the delivery vehicles Job task Industry 産業 load and unload of luggage Firm size 企業規模 8 : From 500 to 999	候補職業 563 a transportation clerk 585 transportation laborers 586 shipping/sorting clerks 594 automobile drivers 568 postal/communication clerks Employment & Job title 雇用形態 2 : Regular employee 1 : No managerial post Attribute (Education) 学歴 9 : Junior high school
---	---

The NANACO system is used for the JGSS (Japanese General Social Surveys) and the SSM (Social Stratification and Social Mobility) survey.

6

Existing Methods (for Binary Classifier)

- **Platt's Method (Sigmoid function)**

$$P(f) = 1 / (1 + \exp(Af + B))$$

- **Zadrozny's Binning Method**

0.1	0.3	0.4	0.5	0.7	0.9	0.9
-----	-----	-----	-----	-----	-----	-----

Accuracy for each bin

- **Isotonic Regression Method (PAV algorithm)**

-2	-1.5	-1.3	-0.5	0	0.2	0.5	0.6	0.8	0.9	score
0	0	0	0	1	0	0	1	0	1	Status(acc uracy)
0	0	0	0	0.3	0.3	0.3	0.5	0.5	1	Accuracy

Expansion by dividing a multiclass classifier into binary classifiers

7

What is the Problem in Multiclass Classification?

- **The relationship among the scores**

the 1st class's score > the 2nd class's score >
the 3rd class's score > ... > the nth class's score

- The 1st class is determined not by the absolute value of the score, but by the **relative position among the scores**.

	Example 1	Example 2
the 1 st class's score	1.5 <i>large</i>	0.1 <i>small</i>
the 2 nd class's score	1.4 <i>large</i>	-1.5 <i>small</i>
the status of the 1 st class	incorrect	correct

8

For Effective Estimation

- Does **class membership probability** for the 1st class depend not only on the 1st class's score but also on **other classes' scores** ?
- It would be better to use not only the 1st class's score, but also **other classes' scores**.

9

Table of Contents

1. Motivation
2. **Proposed Method**
 - a. A Method Using an Accuracy Table
 - b. A Method Applying a Logistic Regression
3. Experiments
4. Conclusion

10

Proposed Method

- Using **multiple classification scores**
- As a method for estimating class membership probabilities
 - a. (indirectly) Using "an accuracy table"
 - b. (directly) Applying a logistic regression

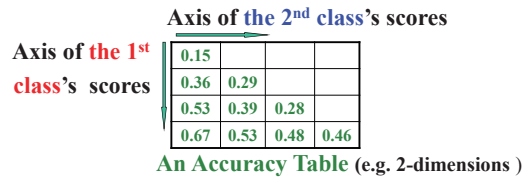
11


Table of Contents

1. Motivation
2. **Proposed Method**
 - a. **A Method Using an Accuracy Table**
 - b. A Method Applying a Logistic Regression
3. Experiments
4. **Conclusions**

12

A Method Using an Accuracy Table



- Both **Binning Method** and **Isotonic Regression Method** are **difficult** to be extended for multi-dimensions.  These methods are not easy to **sort** all samples according to some criteria.

13

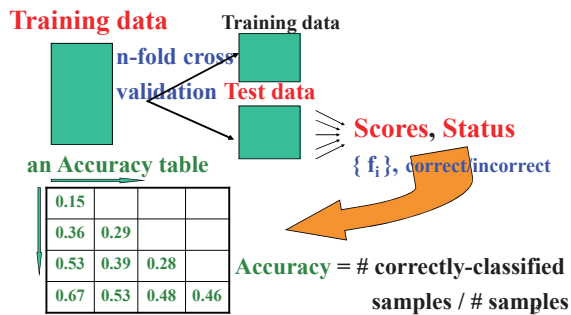
Process

Using **multiple scores**

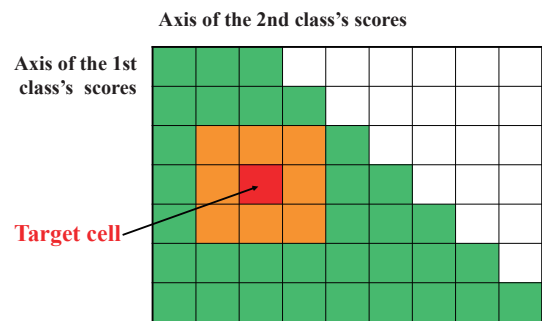
- STEP 1** Create cells for an accuracy table.
- STEP 2** Smooth accuracies.
- STEP 3** Estimate class membership probability for an evaluation sample.

14

STEP 1 Create Cells for an Accuracy Table



STEP 2 Smooth Accuracies



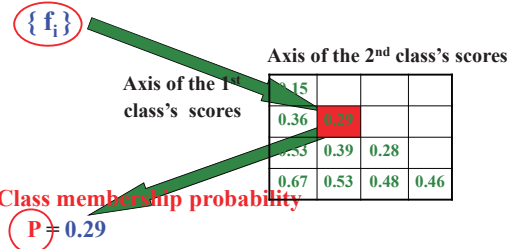
Smoothing Methods

- Using only a **target cell**
 - Laplace's law (Lap) $P_{Lap}(f) = (N_p(c(f)) + 1) / (N(c(f)) + 2)$
 - Lidstone method (Lid) $P_{Lid}(f) = (N_p(c(f)) + \delta) / (N(c(f)) + 2\delta)$
- Using not only a **target cell** but also **surrounding cells**
 - moving average method (MA) $P_{MA}(f) = (N_p(c(f)) / N(c(f)) + \sum_{s \in Nb(c(f))} N_p(s) / N(s)) / n$
 - Median Method (Median) $P_{Median}(f) = \text{median}_{s \in Nb(c(f))} \{N_p(c(f)) / N(c(f)), \{N_p(s) / N(s)\}\}$
 - moving average with coverage method (MA_cov) $P_{MA_cov}(f) = (N_p(c(f)) / N(c(f)) C(c(f)) + \sum_{s \in Nb(c(f))} (N_p(s) / N(s)) C(s)) / (C(c(f)) + \sum_{s \in Nb(c(f))} C(s))$

17

STEP 3 Estimate class membership probability for an evaluation sample

Scores of an evaluation sample



18

Table of Contents

1. Motivation
2. Proposed Method
 - a. A Method Using an Accuracy Table
 - b. A Method Applying a Logistic Regression
3. Experiments
4. Conclusion

19

A Method Applying a Logistic Regression

Formula of a Logistic Regression

$$P(f_1, \dots, f_n) = 1 / (1 + \exp(\sum A_i f_i + B))$$

$\{A_i\}$, B : parameter

f_i : the i^{th} class's score

20

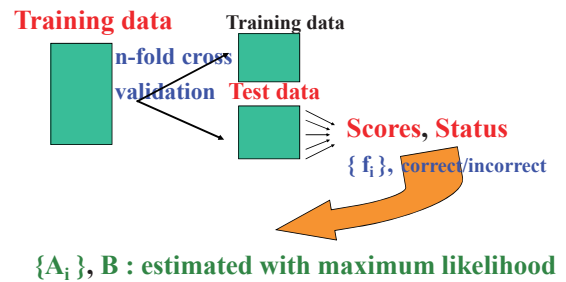
Process

Using **multiple scores**

- STEP 1** Estimate parameter with maximum likelihood method.
- STEP 2** Estimate class membership probability for an evaluation sample.

21

STEP 1 Estimate parameter with maximum likelihood method



22

STEP 2 Estimate class membership probability for an evaluation sample.

Scores of an evaluation sample

$\{f_i\}$

Class membership probability

$$P(f_1, \dots, f_n) = 1 / (1 + \exp(\sum A_i f_i + B))$$

23

Table of Contents

1. Motivation
2. Proposed Method
 - a. A Method Using an Accuracy Table
 - b. A Method Applying a Logistic Regression
3. Experiments
4. Conclusions

24

The Purpose of Experiments

- **Experiment 1**
 - Evaluation of various methods including the proposed methods
- **Experiment 2**
 - Evaluation of the best method

25

Experimental Setting

- **Classifier**
 - **one-versus-rest** method to extend **SVMs** to a multiclass classifier
 - A linear kernel
 - Soft margin parameter $C = 0.6$
 - Features
 - e.g. the JGSS dataset (on the next slide)
 - words in responses to “job task”
 - words in responses to “industry”
 - responses to “employment status” and “job title”
 - **Naïve Bayes classifier**

26

DataSet

- **The JGSS dataset (23,838 samples)**
 - Japanese survey data (open-ended)
 - The number of classes is nearly 200
 - Training data : old data (JGSS-2000, -2001, -2002)
 - Test data : new data (JGSS-2003)
- **The 20 Newsgroups dataset (18,828 samples)**
 - English newspaper articles
 - The number of classes is 20
 - 5-fold cross validation

27

Cell Intervals for an Accuracy Table

- **Cell intervals**
0.05 0.1 0.2 0.3 0.5 etc.

The relationship between **cell intervals** and the **number of cells**

Cell Intervals	0.05	0.1	0.2	0.3	0.5
# cells (the 1 st class's score used)	60	30	16	12	7

28

Evaluation Metrics

- In experiment 1
 - **Negative log-likelihood** a Loss function
$$L = \sum (-y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

y_i : status of an evaluation sample (correct:1 incorrect:0)
 p_i : predicted class membership probability of an evaluation sample

When L is lower, the method is better.
- In experiment 2
 - **Reliability diagram**
the predicted values vs. the true values
 - **ROC (receiver operating characteristic) curve**
FPF (false positive fraction) vs. TPF (true positive fraction)
 - **Ability to detect misclassified samples**

29

The Proposed Method for Creating Cells

Negative Log-likelihood in the best case in each method

Classifier	SVMs	SVMs	Naïve Bayes classifier
DataSet	JGSS dataset	20 Newsgroups dataset	20 Newsgroups dataset
Equal intervals	2369.3 (# cells=30)	1472.3 (# cells=30)	1679.8 (# cells=16)
Equal samples	2678.3 (# cells=12)	1572.9 (# cells=12)	1671.0 (# cells=12)

30

Experiment 1 (1/2)

Negative Log-likelihood (SVMs the JGSS dataset)

Cell Intervals	Used Scores	No Smoothing	Lap	Lid	MA	Median	MA_cov	Logistic regression
0.1	rank1 rank1 & rank2	2309.3 -	2368.9 2356.8	2368.9 2355.8	2367.5 2245.8	2372.6 -	2364.7 2232.7	2367.6 2246.9
0.2	rank1 rank1 & rank2	2371.3 -	2371.0 2252.7	2370.3 2254.7	2369.3 2240.6	2370.0 2341.8	2369.3 2235.0	2367.6 2246.9
0.5	rank1 rank1 & rank2	2381.9 2265.8	2381.8 2265.6	2381.6 2265.7	2395.9 2327.5	2396.4 2298.8	2409.9 2320.6	2367.6 2246.9 ₃₁

Experiment 1 (2/2)

Negative Log-likelihood (SVMs the 20 Newsgroups dataset)

Cell Intervals	Used Scores	No Smoothing	Lap	Lid	MA	Median	MA_cov	Logistic regression
0.1	rank1 rank1 & rank2	1472.3 -	1472.4 1390.2	1472.2 1388.3	1468.1 1362.3	1469.6 -	1467.4 1360.3	1482.3 1386.6
0.2	rank1 rank1 & rank2	1472.5 -	1472.7 1365.4	1472.5 1366.9	1474.4 1374.9	1473.3 -	1482.7 1377.7	1482.3 1386.6
0.5	rank1 rank1 & rank2	1487.4 1388.1	1487.5 1387.7	1487.4 1387.8	1503.9 1447.2	1497.0 1408.7	1537.9 1479.4	1482.3 1386.6 ₃₂

Negative Log-likelihood with SVMs on Both Datasets

- A method using an accuracy table
rank1 & rank2 < rank1 & rank2 & rank3 << rank1
 - A method applying a logistic regression
rank1 & rank2 & rank3 < rank1 & rank2 << rank1
- ↓
- Using multiple scores was much effective in SVMs
 - The method using an accuracy table (cell intervals = 0.1 and a smoothing method = MA_cov) was the best of all cases.
 - A method applying a logistic regression was stable.

33

Negative Log-likelihood with Naïve Bayes classifier on the 20 Newsgroups dataset

In the case of the method using an accuracy table

# Cells (the 1st class's score used)	Used Scores	No smoothing	Lap	MA	Median	MA_cov
30	rank1 rank1 & rank2	- -	1680.6 1439.7	1670.1 1409.8	1668.4 -	1675.0 1415.3
16	rank1 rank1 & rank2	1680.2 -	1679.8 1428.1	1679.6 1515.5	1675.8 -	1696.2 1536.2
7	rank1 rank1 & rank2	1697.2 -	1697.2 1474.8	1712.0 1626.3	1713.6 1644.8	1732.8 1664.1 ₃₄

The best method

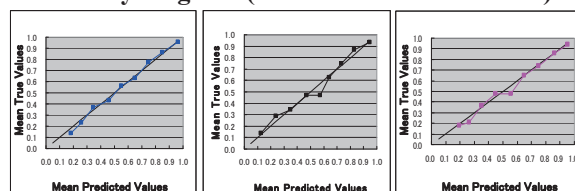
The method using both the 1st class's score and the 2nd class's score

- A Method using an accuracy table
- A Smoothing method : MA_cov
- Cell intervals : 0.1 (# cells : 30)

35

Experiment 2

Reliability diagram (SVMs the JGSS dataset)



The best method

Using the 1st class's score & the 2nd class's score

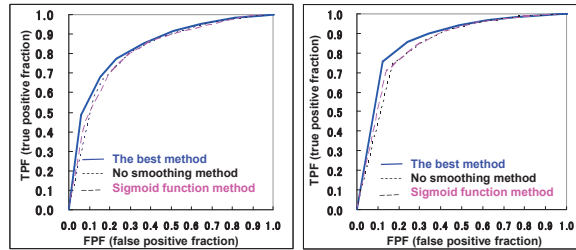
No smoothing method

Using the 1st class's score Binning method (# bins=30)

Sigmoid function method

Using the 1st class's score Platt's method₃₆

ROC Curve with SVMs

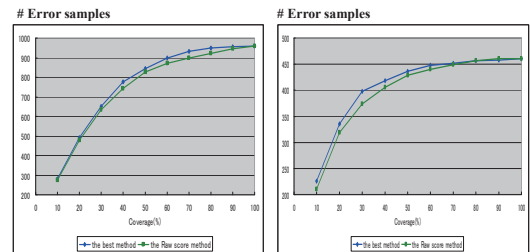


The JGSS dataset

The 20 Newsgroups dataset

37

Ability to detect misclassified samples with SVMs



The JGSS dataset

The 20 Newsgroups dataset

38

Table of Contents

1. Motivation
2. Proposed Method
 - a. A Method Using an Accuracy Table
 - b. A Method Applying a Logistic Regression
3. Experiments
4. Conclusions

39

Conclusions

- To estimate class membership probabilities, we proposed using **multiple classification scores** outputted by classifiers.
- As a method for estimating class membership probabilities
 - Generating an accuracy table with smoothing methods such as the moving average method or the moving average with coverage method
 - Applying a logistic regression
- We empirically showed that the use of **multiple classification scores** was much effective in both methods.
- We also showed that the proposed smoothing method for the accuracy table works quite well, and that the method applying a logistic regression is more stable.

40

Future Work

- We would like to find an effective method for estimating class membership probabilities for **any class** in multiclass classification.

41

Thank you !

42

自由回答分類としての 産業・職業分類 自動コーディングの開発と活用

敬愛大学国際学部
高橋 和子

1

於:統計センター 2007年12月12日

発表の順序

1. はじめに
2. 職業コーディング
3. 機械学習とルールベース手法の組み合わせによる自由回答の自動分類
4. 複数の分類スコアを用いたクラス所属確率の推定
5. コーダの分類作業を支援する実用システム(NANACOシステム)の開発
6. おわりに

2

1.1 背景と目的 (1/2)

● 研究の背景

社会調査における代表的な回答形式

- 選択回答 …… 分析者の枠組みによる 構造化データ(コード)
調査票に提示された選択肢のみ回答
- 自由回答 …… 回答者の枠組みによる非構造化データ(テキスト)
自由に記述できるため選択回答より豊かな情報
統計処理のためにコード変換(コーディング)が必要

[コーディングにおける問題点]

- 作業内容が複雑であり、作業量が多い
- コーディング結果の妥当性と信頼性が保証されない(一貫性がない)

自由回答の利点を生かすためにコーディング(分類)方法についての検討を行い、社会調査技術の一つとして確立したい

3

背景と目的 (2/2)

● 研究の目的

機械学習を用いて自由回答を高精度に自動分類する方法および分類結果のクラス所属確率を高精度に推定する方法を提案

[想定する自由回答分類]

- 分類カテゴリが定義済み
- トピックに注目
- 分類の単位は回答全体
- 分類カテゴリが多数

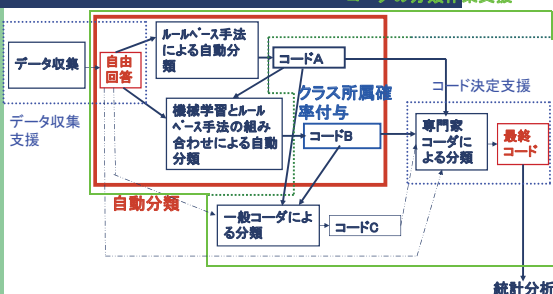
代表的な自由回答分類である

職業コーディングを例として用いる

4

自由回答の処理過程における研究の位置づけ ー職業コーディングの場合ー

コードの分類作業支援



5

2.1 職業コーディングとは

社会調査における統計分析のために、自由回答で収集される職業データを総合的に判断して職業コード(1つ)に変換する作業

[例]

● 職業データ(SSM調査の場合)

- Q1-1 仕事の内容 : 配車等を手配
- Q1-2 従事先事業の種類 : 荷物のつみおろし業務他
- Q1-3 従事上の地位 : 2(常時雇用の一般従事者)
- Q1-4 役職 : 1(役職なし)
- Q1-5 従事先事業の規模 : 8(500~999人)

↓ これまで*は人手による分類

- 職業コード : 563(運輸事務員) ← SSM職業コード(約200)

他にも国際標準コードISCO(International Standard Classification of Occupations)コード(約400種類)あり

6

2.2 職業コーディングにおける問題

- 一般の自由回答分類における問題
+
- 職業コーディング特有の問題
 - カテゴリが非常に多い(数百個)
 - 複数の質問に対する回答から総合的に判断する
 - 調査によっては本人現職以外にも多数の職業を収集する

↓
一般の自由回答分類より困難なタスク
自動分類を検討する必要性が高い

7

2.3 海外における自動分類の状況*

- オーストラリア
Precision Data パターンマッチ (シソーラスベース)
- 米国
 - AIOCS ルールベース (正解率47%) 1990年国勢調査に適用
 - PACE Memory Based Reasoning (MBR) (正解率60%)
超高性能なコンピュータが必要
- フランス
SICORE パターンマッチ (2-gramの木構造)
- カナダ
ACTR 一般的な自動コーディングシステム(1語または2語のみ対応)で代用

職業コーディングをカテゴリーのある分類タスク (classification)として捉え、自然言語処理や機械学習による方法を検討

8

3.1 自動職業コーディングの目標とルールベース手法における問題点

自由回答の高精度な自動分類のために、機械学習手法の一つであるサポートベクターマシン(SVM)によるアプローチおよび、SVMと先に開発したルールベース手法との組み合わせ方を検討し有効な方法を提案する

- 正解：職業コーディングにおいて専門家コードが最終的に決定したコード
- 正解率＝正解事例数／コードを付与した事例数
- 正解率の目標値：約75%
一般コードの正解率(約68.8～80.0%)の平均

まず、ルールベース手法を開発し5つの調査に適用

職業を動作(述語)に注目し、格フレームの概念により表現
決定できた事例約80% × 決定した事例の約80%は正しかった
→ 正解率(約65～70%)が目標値に達しなかった

9

3.2 ルールベース手法

- 文を形態素解析により語に切り分け、不要語(等、など、...)を削除しておく
- 職業の定義を格フレームの形式により記述したルール(ルールα)および職業データを総合的に判断するルール(ルールβ)を生成しておく
- 自由回答を格フレームの形式により表現する
- 自由回答からルールαのセットを検索し、仮の職業コードを決定(該当するルールαがなければ未定コード「999」を付ける)
 - 格フレームにおける述語や名詞をそれぞれ述語シソーラス、名詞シソーラスにより拡張
- 仮の職業コードと職業データからルールβのセットを検索し職業コードを決定(該当するルールβがなければ仮の職業コードで決定)

10

ルールベース手法の手順

- STEP1 『職業定義辞書』に記述された内容を格フレームの形式で表現した三つ組みと職業コードによるルールセットとして作成しておく
ルールα：<述語, 格, 名詞> → <職業コード>
3,524個
総合的に判断するための知識もルールセットとして作成しておく
ルールβ：<職業コード, 従業上の地位, 役職, 従業先事業の規模>
27個 → <職業コード>
- STEP2 職業データ*の自由回答を格フレームによる三つ組み
<述語, 格, 名詞>で表現
- STEP3 回答中の「述語」や「名詞」をシソーラスにより拡張
- STEP4 回答からルールセット(ルールα)を検索し、仮の職業コードを付与
- STEP5 自由回答以外の職業データも参照し、ルールセット(ルールβ)を検索し、最終的な職業コードを付与

11

ルールベース手法の例

- | | |
|---------------|-----------------|
| Q1-1 仕事の内容 | ： 配車等を手配 |
| Q1-2 従業先事業の種類 | ： 荷物のつみおろし業務他 |
| Q1-3 従業上の地位 | ： 2(常時雇用の一般従事者) |
| Q1-4 役職 | ： 1(役職なし) |
| Q1-5 従業先事業の規模 | ： 8(500～999人) |
- STEP1 ルールセットを作成
- STEP2 「仕事の内容」：<手配 ヲ 配車>
STEP3 述語 手配 を拡張し述語コード(385 10)を付ける
STEP4 ルールα <385 10, ヲ, 配車> → <563>
STEP5 ルールβ 今回は該当するものなし
- ↓
「職業コード」： 563 (運輸事務員)

12

ルールベース手法の利用状況と性能

- 利用された主な調査*
 - JGSS (日本版General Social Surveys) 予備調査 (1回)、2000~2003年調査 (4回)
 - SSM (Social Stratification and Social Mobility) 調査 2005年調査のための予備調査 (2003SSM予備調査)
 - 他4つの調査
 - 正解率 約65~70% < 75% (目標値)
全サンプルの約80%しか決定できなかったが、
決定した事例の約80~85%は正しかった
- 【問題点】
- すべての知識をルール化するのは困難である
 - シソーラスやルールセットのメンテナンスが継続的に必要である
 - 格フレームの形式で表現できない回答に対応できない

13

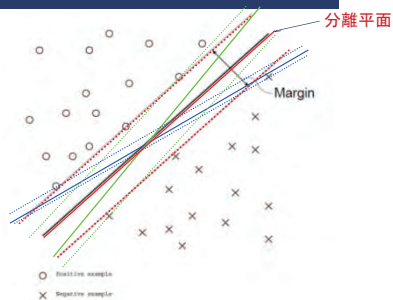
3.3 機械学習による方法

サポートベクターマシン(SVM)の優位性

- SVMは多くの文献で文書分類における分類性能の高さが示されている ← マージン最大化戦略により汎化能力が高い [文書分類の場合]
 - 新聞記事 (RWCPコーパス) の分類
決定木学習法と比較
 - 新聞記事 (Reuters collection) や医学論文の要約 (Ohsumedコーパス) の分類
ナイーブベイズ分類器、決定木学習法、k-NN法と比較
Find Similar、ナイーブベイズ分類器、決定木学習法、ベイズネットと比較
 - 同一のコーパスを対象とする既存研究のまとめ
ナイーブベイズ分類器、決定木学習法、k-NN法などと比較

14

マージン最大化戦略の例



15

SVMによる方法の手順

- STEP1 全事例 (職業データ) から抽出した素性により素性辞書を作成しておく
- STEP2 正解のわかっている事例 (職業データ) から抽出した素性と正解 (職業コード) の組により訓練データを生成する
- STEP3 訓練データを分類器*に学習させる
- STEP4 分類器を評価データに適用する

分類器* : one-versus-rest法により多値分類器に拡張

16

SVMによる訓練データの生成例 (STEP2)

- 回答から基本的な素性を抽出
 - 配車, を, 手配 ... 「仕事の内容」
 - 荷物, の, つみおろし, 業務, 他 ... 「従業先事業の種類」
 - 2 (常時雇用の一般従事者 役職なし) ... 「従業上の地位と役職」

・正解
563

563 配車 を 手配 荷物 の つみおろし 業務 他 2

*実際には各素性は素性辞書により素性番号に変換されている

17

SVMによる方法の性能

- データセット : JGSS (日本版General Social Surveys)
 - 訓練データ ... JGSS-2000, -2001 (13,296サンプル)
 - 評価データ ... JGSS-2002 (6,770サンプル)
 - 分類器 : one-versus-rest法により多値分類器に拡張 (ソフトマージンパラメタ C=1.0)
 - カーネル関数 : 線形カーネル
 - 正解率 : 約66% (ルールベース手法) < 約72% < 75% (目標値)
ルールベース手法より6%高かったが、目標値には達しなかった
決定された事例においてはルールベース手法より8%低かった
- 【利点】
- 人手でルールを作成する必要なし
 - シソーラスやルールセットのメンテナンスが不要
 - 格フレームの形式で表現できない回答も処理可能

18

SVMによる効果的な方法の検討

- 実験設定の変更
JGSS-2001 (6,448サンプル)を用いた10分割交差検定
 - 素性の組み合わせ
例 2-gram, 3-gram をとる
「従業先事業の規模」も利用
単語を「原形」ではなく「読み」にする
 - 素性選択
例 Information Gainの利用
 - カーネル関数の次元
例 2,3次元に上げる
- 効果なし
- SVMとルールベース手法の組み合わせを検討

19

3.4 機械学習とルールベース手法との組み合わせによる方法

SVMとルールベース手法の組み合わせ方

- SVMの基本素性に、ルールベース手法が出力した職業コードを追加 (add-code)
- SVMの基本素性に、ルールベース手法でマッチしたルールを追加 (add-rule)
- SVMの基本素性に、ルールベース手法が出力した職業コードとルールベース手法でマッチしたルール(2種類)を追加 (add-code-rule)
- ルールベース手法が職業コードを決定できない場合に、SVM(単独)の結果を利用 (seq)

20

add-codeとadd-ruleの場合

- add-rule
SVMの素性に追加
素性1、...、素性n+1、素性n+2
 - add-code
SVMの素性に追加
素性1、...、素性n+1
- ルールベース手法
ルールα100
ルールβ10
職業コード○○○
- * スタッキングの一種であると考えられる

21

seqの場合

	SVMの出力	ルールベース手法の出力	用いる結果
例1	○○○	1個(563)	→ 563
例2	△△△	未定(999)	→ △△△
例3	×××	複数個(501、503)	→ ×××

* アンサンブル学習の一種であると考えられる

22

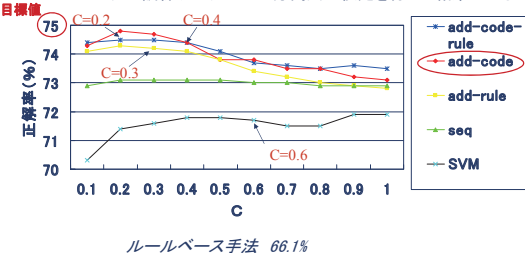
3.5 実験

- 実験1
機械学習とルールベース手法の有効な組み合わせ方を調査
 - 実験2
訓練データの量と正解率の関係を調査
- * 実験設定はSVM単独の場合と同様
ソフトマージンパラメタCの値
実験1 0.1~1.0まで10通りに変化
実験2 1.0に固定

23

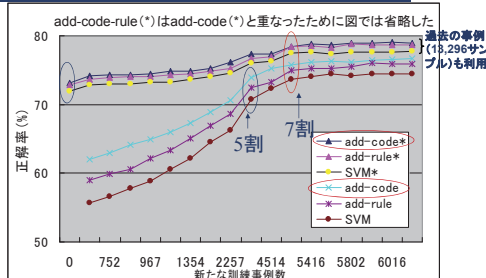
実験1の結果 各手法の正解率とCのチューニング

Cのチューニングは訓練データ内で10分割交差検定を行った結果による目標値



24

実験2の結果 訓練事例数と正解率の関係



25

3.6 提案手法の適用結果と評価 正解率 (1/2)

add-code JGSSにおける ROCCO システムと提案手法による比較 (単位: %)

システム名	ROCCO	ROCCO	ROCCO	ROCCO	提案手法
調査名	JGSS-2000	JGSS-2001	JGSS-2002	JGSS-2003	JGSS-2003
精度	80.0	80.5	79.4	79.8	80.7
再現率	65.8	66.6	66.7	64.1	80.7

74.4(C=0.4)

訓練事例の増加 (13,296サンプル→20,066サンプル)

訓練データと評価データの調査主体が異なる場合 72.4% (< 目標値 75%)
(2005SSM調査)

訓練事例の増加 (20,066サンプル→34,521サンプル)

26

正解率 (1/2)

- 訓練データと評価データの調査主体が異なる場合も第1位から第5位までの正解率は約85%
→ コーダが判断する際の参考にできる
(→ コーダの分類作業支援に役立ちそう)
- 処理時間 (add-code の場合)
約0.21秒/サンプル
(AMD Athlon 64 X2 Dual Core Processor 4400+ × 2 processor)

27

3.7 まとめ

- 自由回答を高精度に自動分類するために、職業コーディングを例として用い、機械学習 (SVM) を適用する方法を検討した
- SVMによる方法はルールベース手法より有効であったが、SVMによる方法とルールベース手法を組み合わせる手法 (4つ) はさらに有効であった
add-code-rule ≧ add-code > add-rule > seq > SVM単独
- 特に、素性としてルールベース手法が出力した分類結果を追加する方法 (add-code) は有効で目標であるコーダの正解率 (75%) をほぼ達した
- ただし、ソフトマージンパラメータCのチューニングが必要
- 過去の事例の利用は有効であり、新たな事例をフィードバックすることはさらに有効であった
- 提案手法は実際の調査に適用された結果、正解率は、訓練データと評価データの調査主体が同一の場合81%、異なる場合72%であった
- 訓練データと評価データの調査主体が異なる場合はフィードバックが有効であると思われるが、今後実験により確認する必要がある

28

4.1 クラス所属確率の有用性と従来手法における問題点

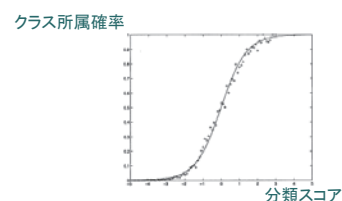
多値分類における任意のクラスについてのクラス所属確率を複数の分類スコアを用いて推定する方法 (直接的な方法と間接的な方法) を提案する

- 自由回答の自動分類結果が利用できる場合に、事例が分類器により予測されたクラスに属する確率 (クラス所属確率) がわかれば便利である
- 文書分類における多くのアプリケーションにおいてもクラス所属確率は有用で、正確な推定値が必要とされる
- 既存の推定方法*はいずれも2値分類を想定するため、推定したいクラスの分類スコアのみを用いている
- しかし、第1位のクラスは、分類スコアの絶対的な大きさではなく相対的な大きさにより決定される
→ 第1位のクラスのクラス所属確率は、第1位のクラスの分類スコアだけでなく他のクラスの分類スコアにも依存する
→ 第1位のクラスだけでなく他のクラスの分類スコアも用いた方がより正確な推定ができると考えられる

29

〔既存の研究〕 クラス所属確率の推定

- Plattの方法 (シグモイド関数) (直接推定) (Platt, 1999)
 $P(f) = 1 / (1 + \exp(Af + B))$ f: 分類スコア A, B: パラメータ



30

〔既存の研究〕 クラス所属確率の推定

- Binning (間接推定) (Zadrozny and Elkan, 2001)
等サンプルのビン

0.1	0.3	0.4	0.5	0.5	0.7	0.9	正解率
-----	-----	-----	-----	-----	-----	-----	-----

- Isotonic regression (間接推定) (Zadrozny and Elkan, 2002)
- PAV (Pool Adjacent Violators) アルゴリズム

-2	-1.5	-1.3	-0.5	0	0.2	0.5	0.6	0.8	0.9	分類スコア
0	0	0	0	1	0	0	1	0	1	正解/不正解
0	0	0	0	0.3	0.3	0.3	0.5	0.5	1	正解率

31 BinningとIsotonic regressionは事例を分類スコアによりソートする必要あり

4.2 第1位のクラスについてのクラス所属確率推定 4.2.1 提案手法

- 複数の分類スコアを用いてクラス所属確率を推定
- 推定の方法
 - [パラメトリックな方法]
ロジスティック回帰を適用する方法 (直接推定)
 - [ノンパラメトリックな方法]
正解率表を利用する方法 (間接推定)

32

ロジスティック回帰を適用する方法

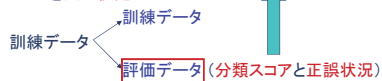
$P = 1 / (1 + \exp(-\sum A_i f_i + B))$ により推定

f_i : 分類スコア A_i, B : パラメタ

[手順]

STEP 1 ロジスティック回帰式のパラメタ A_i, B を最尤推定

訓練データを交差検定



STEP 2 評価事例の分類スコア f_i をロジスティック回帰式に代入して
クラス所属確率を推定

33

正解率表を利用する方法

正解率表

分類スコアを2つ用いた場合の正解率表の例 (2次元)
分類スコア軸 (第2位のクラス)

分類スコア軸 (第1位のクラス)	0.15			
	0.36	0.29		
	0.53	0.39	0.28	
	0.67	0.53	0.48	0.46

用いる分類スコアの数 は任意 (r個の場合はr次元)

* 従来の方法 (BinningやIsotonic Regression) は多次元への
拡張が困難

34

正解率表を利用する方法の手順

STEP 1 訓練データの分類スコアを軸として等間隔に区切り正解率表作成
のためのセルを作成

STEP 2 セルごとに正解率 (=セル内正解事例/セル内全事例) を計算

STEP 3 正解率を平滑化* (次スライド)

STEP 4 評価事例の分類スコアにより正解率表の該当するセルの
正解率をクラス所属確率として間接的に推定

35

正解率表の平滑化手法

- 注目するセルの情報だけを用いる方法

- ラプラス法 $P_{Lap}(f) = (N_p(c(f)) + 1) / (N(c(f)) + 2)$

- リッドストーン法 $P_{Lid}(f) = (N_p(c(f)) + \delta) / (N(c(f)) + 2\delta)$

- 周囲のセルの情報も用いる方法

- 移動平均法 $P_{MA}(f) = (N_p(c(f))/N(c(f)) + \sum_{s \in Nb(c(f))} N_p(s)/N(s)) / n$

- メディアン法 $P_{Median}(f) = median_{s \in Nb(c(f))} \{N_p(c(f))/N(c(f)), \{N_p(s)/N(s)\}\}$

- カバレッジを重みとする移動平均法

$$P_{MA_cov}(f) = (N_p(c(f))/N(c(f))C(c(f)) + \sum_{s \in Nb(c(f))} (N_p(s)/N(s))C(s)) / (C(c(f)) + \sum_{s \in Nb(c(f))} C(s))$$

36

4.2.2 実験

- 実験1
 - 第1位のクラスのクラス所属確率を推定するための有効な方法を調査
- 実験2
 - 実験1で最良であった推定方法の評価

37

実験設定 (1/2)

- 分類器
 - 多値分類器に拡張したSVM (ソフトマージンパラメタ $C=0.6$)
 - ナイーブベイズ分類器
- データセット (2種類)
 - JGSSデータセット ... 日本語自由回答
 - クラス数 約200
 - サンプル数 23,838 (JGSS-2000, -2001, -2002, -2003)
 - 20Newsgroupsデータセット ... 英文ネットニュース記事
 - クラス数 20
 - サンプル数 18,828 (5分割交差検定)

38

実験設定 (2/2)

- ロジスティック回帰式パラメタの最尤推定
 - 訓練データ内における5分割交差検定
- 正解率表のセル幅 (分類スコアの区間幅) の設定 (5通り)
 - 0.05 0.1 0.2 0.3 0.5
- 評価尺度
 - クロスエントロピー $L = \sum (y_i \log(p_i) + (1-y_i) \log(1-p_i)) / N$ 小さいほどよい
 - Reliability Diagram 予測値 対 実測値 重なっているほどよい
 - ROC (receiver operating characteristic) 曲線 左上にあるほどよい
 - TPF (True Positive Fraction) = 正解事例数 / 全正解数
 - 対 FPF (False Positive Fraction) = 不正解事例数 / 全不正解数
 - AUC (Area Under the Curve) 大きいほどよい
 - 誤分類検出能力 カバレッジが低いときに多いほどよい

39

実験1 結果のまとめ (クロスエントロピーによる評価)

- ロジスティック回帰による方法 (SVM)
 - 第1位から第3位までの分類スコア (3個) を用いることが有効
 - 用いる分類スコアの数が多い (ロジスティック回帰式のパラメタ数が多い) ほど正確な推定が行われた
 - 結果が安定
- 正解率表を利用する方法 (SVM ナイーブベイズ分類器)
 - 第1位と第2位の分類スコア (2個) を用いることが有効
 - 用いる分類スコアの数が多すぎると事例が少ない (または含まれない) セルが増え、正解率表の精度が低下した
 - ロジスティック回帰による方法も含めすべての方法の中で最もよかったのは、最適な正解率表 (セル幅0.1 カバレッジを重みとする移動平均法による平滑化手法を適用) を利用する方法であった

40

実験1の結果 SVM クロスエントロピー* (JGSS データセット 3,722サンプル)

セル幅	用いた分類スコア	平滑化なし	ラプラス法	リッドストーン法	移動平均法	メディアン法	カバレッジ付き移動平均法	ロジスティック回帰*
0.1	第1位	2309.3	2368.9	2368.9	2367.5	2372.6	2364.7	2367.6
	第1位と第2位	-	2356.8	2355.8	2245.8	-	2232.7	2246.9
0.2	第1位	2371.3	2371.0	2370.3	2369.3	2370.0	2369.3	2367.6
	第1位と第2位	-	2252.7	2254.7	2240.6	2341.8	2235.0	2246.9
0.5	第1位	2381.9	2381.8	2381.6	2395.9	2396.4	2409.9	2367.6
	第1位と第2位	2265.8	2265.6	2265.7	2327.5	2298.8	2320.6	2246.9

正解率表を利用する方法はいずれも「第1位と第2位 < 第1位から第3位まで < 第1位のみ」の順ロジスティック回帰* において第1位から第3位まで用いた場合は2232.9

41

実験1の結果 SVM クロスエントロピー* (20newsgroupsデータセット 3,765サンプル)

セル幅	用いた分類スコア	平滑化なし	ラプラス法	リッドストーン法	移動平均法	メディアン法	カバレッジ付き移動平均法	ロジスティック回帰*
0.1	第1位	1472.3	1472.4	1472.2	1468.1	1469.6	1467.4	1482.3
	第1位と第2位	-	1390.2	1388.3	1362.3	-	1360.3	1386.6
0.2	第1位	1472.5	1472.9	1472.5	1474.4	1473.3	1482.7	1482.3
	第1位と第2位	-	1365.4	1386.9	1374.9	-	1377.7	1386.6
0.5	第1位	1487.4	1487.5	1487.4	1503.9	1497.0	1537.9	1482.3
	第1位と第2位	1388.1	1387.8	1387.8	1447.2	1408.7	1479.4	1386.6

正解率表を利用する方法はいずれも「第1位と第2位 < 第1位から第3位まで < 第1位のみ」の順ロジスティック回帰* において第1位から第3位まで用いた場合は1377.5

42

実験1の結果 ナイーブベイズ分類器 クロスエントロピー*(20newsgroupsデータセット 3,765サンプル)

セル数	用いた分類スコア	平滑化なし	ラプラス法	移動平均法	メディアン法	カハレツジ付き移動平均法
30*	第1位	-	1680.6	1670.1	1668.4	1675.0
	第1位と第2位	-	1439.7	1409.8	-	1415.3
16	第1位	1680.2	1679.8	1679.6	1675.8	1696.2
	第1位と第2位	-	1428.1	1515.5	-	1536.2
7	第1位	1697.2	1697.2	1712.0	1713.6	1732.8
	第1位と第2位	-	1474.8	1626.3	1644.8	1664.1

30*: SVMにおいてはセルの区間幅0.1の場合にセル数30個であった

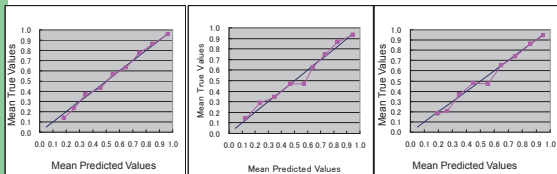
43

実験2 結果のまとめ(最良の方法に対する評価)

- 両方のデータセット
 - 最良の方法は、Reliability Diagramによる評価においてもROC曲線(AUC)による評価においても従来手法より有効であった(SVM)
 - 従来手法 : 第1位の分類スコアのみを用い平滑化なし
第1位の分類スコアのみを用いシグモイド関数適用
- 最良の方法は既存の方法より誤分類の検出能力が高かった(SVM)
- 既存の方法 : 分離平面からの距離を単純に確率にする既存の方法 (Schohn and Cohn, 2000)

44

実験2の結果 最良の方法のReliability Diagramによる評価(SVM JGSSデータセット)



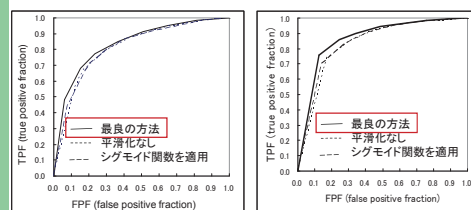
第1位と第2位のクラスの分類スコアを用いる
第1位のクラスの分類スコアのみを用いる
第1位のクラスの分類スコアのみを用いる

最良の方法(最適な正解率表を利用: 区間幅0.1かつカハレツジを重みとする移動平均法による平滑化)
平滑化なし(≒ Binning)
シグモイド関数を適用(≒ Plattの方法)

45

20NewsGroupsデータセットにおいても同様の結果であった

実験2の結果 最良の方法のROC曲線による評価(SVM)

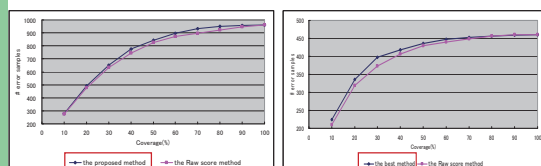


JGSSデータセット

20NewsGroupsデータセット

46

実験2の結果 最良の方法における誤分類の検出能力(SVM)



JGSSデータセット

20NewsGroupsデータセット

the Row score method : 分離平面からの距離を単純に確率値に直す既存の方法 (Schohn and Cohn, 2000)

47

4.3 第2位以下の任意のクラスについてのクラス所属確率推定

第1位のクラスの結果を第2位以下の任意のクラスに拡張(どのクラスの分類スコアを用いるのが有効か?)

[提案手法]

- クラス所属確率を推定したいクラス(注目するクラス)と第1位のクラスの分類スコアを用いる

第1位のクラスの分類スコア > 第2位のクラスの分類スコア > ... > 注目するクラスの分類スコア > ... > 第n位のクラスの分類スコア

- ロジスティック回帰を適用する ← 安定性
第k位のクラスについてのクラス所属確率を推定したい場合

$$P(f_1, f_k) = 1 / (1 + \exp(A_1 f_1 + A_k f_k + B))$$

f_1, f_k : 分類スコア A_1, A_k, B : パラメタ

48

4.3.2 実験

- 実験1
 - 第2位以下の任意のクラスのクラス所属確率を推定するための有効な分類スコアの組み合わせを調査
- 実験2
 - クラス所属確率の推定方法として、ロジスティック回帰を適用する方法の有効性を確認

* 実験設定は5.2の実験と同様

分類器: 多値分類器に拡張したSVM (ソフトマージンパラメタ $C=0.6$)

49

実験1 結果のまとめ

両方のデータセット

- クロスエントロピーによる評価(第2位から第20位までのクラスに注目)
注目するクラス(第k位) & 第1位のクラス \Rightarrow 第1位のクラスから注目するクラス(第k位)までのすべてのクラス $<$ 注目するクラス(第k位) & 直前直後のクラス(第k-1位 & 第k+1位) \Rightarrow 注目するクラス(第k位)のみ
- ROC曲線による評価(第2位から第4位までのクラスに注目)
注目するクラス(第k位) & 第1位のクラス \Rightarrow 第1位のクラスから注目するクラス(第k位)までのすべてのクラス $<$ 注目するクラス(第k位) & 直前直後のクラス(第k-1位 & 第k+1位) \Rightarrow 注目するクラス(第k位)のみ

50

実験2 結果のまとめ

両方のデータセット

- クロスエントロピーによる評価(第2位から第5位までのクラスに注目)
ロジスティック回帰による方法 \Rightarrow 「正解率表(改良版)」を利用する方法 $<$ 「正解率表」を利用する方法
- ROC曲線による評価(JGSSデータセット 第2位のクラスに注目)
ロジスティック回帰による方法 \Rightarrow 「正解率表(改良版)」を利用する方法 $<$ 「正解率表」を利用する方法

ただし、注目するクラスごとに「正解率表(改良版)」を作成するのは困難

51

4.4 まとめ

- 多値分類における任意のクラスのクラス所属確率について、複数の分類スコア(推定したいクラスと第1位のクラスの分類スコア)を用いてロジスティック回帰により高精度に推定する方法を提案した
- クラス所属確率を推定する別の方法として、分類スコアを軸として等区間に区切って正解率を計算しておく「正解率表」を利用する方法も提案した
- 「正解率表」を利用する方法では、周囲のセルの正解率も利用した平滑化手法(移動平均法など)が有効であった
- 複数の分類スコアを用いる方法は、いずれの推定方法も日本語自由回答と英文ネットニュース記事の2つのデータセットにおいて有効性を示した
- 特にロジスティック回帰による方法は安定してよい結果を示し、「正解率表」を利用する方法は、区間幅を最適に設定すればロジスティック回帰による方法を上回る場合もあったが、最適な正解率表の作成は困難であった

52

5.1 システムの目的と構成

産業・職業コーディングにおいて自動分類システムが利用され始めたが、結果がそのまま採用されるわけではなく、一般コーダの作業は依然として残る。

また、最終的な決定も人間により行われる

人間による判断(分類作業)を支援するために必要な情報をわかりやすく提示するシステムが必要ではないか

[目的]

コーダが短時間で正確な作業を一貫性を持って行えるように、自動分類結果を、コーディングに必要な他の情報とともにわかりやすく提示するシステム(NANACOシステム)を開発する

53

システムの構成図

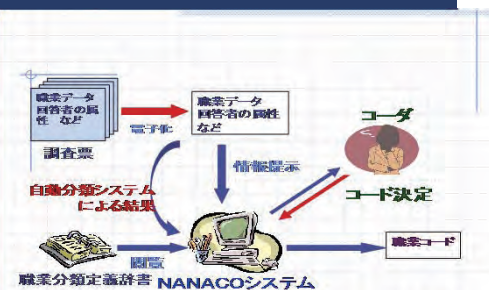
自動分類

クラス所属確率の推定

(狭義の)
NANACOシステム
(ユーザインタフェース)

54

5.2 自由回答分類におけるシステムの位置づけ



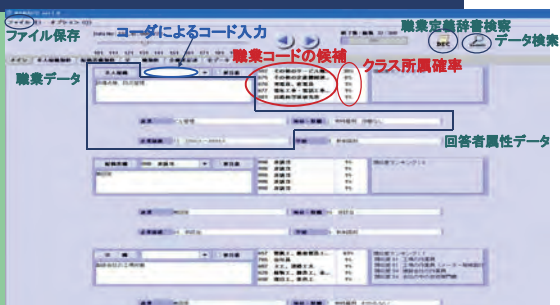
55

5.3 システムにおける主な機能と処理概要

- 主な機能
 - 自動分類結果の上位にランキングされたクラスを候補として確信度（クラス所属確率）付きで提示
 - コーディングの対象となるデータおよび回答者の基本的な属性データの提示
 - カテゴリの定義内容を記述した辞書の検索および閲覧表示
 - 関連データの検索
 - コードにより付与されたコードのファイル保存
- 処理概要
 - ファイルの読み込み
 - 情報提示
 - 結果ファイルの出力

56

情報提示の例 (NANACOシステム作業画面例)



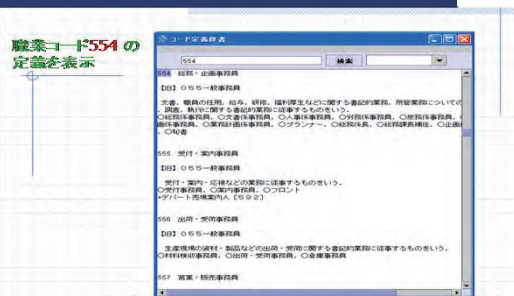
57

職業コード定義の閲覧例



58

『職業定義辞書』の表示例



59

データの検索結果例



60

5.4 システムの評価

● NANACOシステムの適用実績(5つの調査)

- ◆ JGSS-2003
 - 3,663サンプル×3(本人現職、配偶者職、父職)
- ◆ 「暮らしと健康に関する全国調査」(2003年実施)
 - 1,910サンプル(全員有職者)×1(本人現職)
- ◆ 2003SSM調査(2005SSM調査の予備調査)
 - 767サンプル×2(本人現職、本人副職)
- ◆ 2005SSM調査
 - 5,764サンプル×6以上(本人現職、本人初職、本人職歴全部、配偶者職、父職(本人15才時)、父職(主なもの)、母職)
- ◆ 2005SSM若年層調査
 - 1,191サンプル×5(本人現職、本人初職、本人3年後職、配偶者職、父職)

● 評価項目

- コーダの正解率
- コーダの作業時間
- コーダによる評価

61

5.4.1 コーダの正解率

JGSS-2001
経験者

		コードの正解率(単位: %)				
NANACO利用の有無	調査名	本人現職	本人最後職	本人初職	配偶者職	父職
有	JGSS-2003	93.2	収集されず	収集されず	95.0	96.7
無*	(平均)	88.9	90.1	88.7	85.9	89.4
無*	JGSS-2002	95.7	94.3	92.3	91.9	94.3
無*	JGSS-2001	82.1	85.8	85.1	79.8	84.5
無	JGSS-2000	78.1	72.1	79.0	68.8	70.7

無*: ルールベース手法(ROCCOシステム)による分類結果を利用

NANACOシステムの有効性 ← 精度の高い候補を提示 & 情報の有効活用

62

5.4.2 作業時間

計測した調査

- JGSS-2003
- 「暮らしと健康に関する全国調査」(2003年調査)

● JGSS-2003(有職者 約6割)の場合

NANACOシステム利用 平均0.6分/サンプル

● 「暮らしと健康に関する全国調査」(すべて有職者)の場合

{ NANACOシステム利用グループ(4人) 平均1.5分/サンプル
ROCCOシステム利用グループ(4人) 平均5.3分/サンプル

63

単位人時間あたりの処理サンプル数

NANACOシステム利用の有無による処理サンプル数の比較

	第1日目午後	第2日目午前	第2日目午後	平均
NANACO利用の有無	処理サンプル数/人時間	処理サンプル数/人時間	処理サンプル数/人時間	処理サンプル数/人時間
有	24.4	36.7	50.8	37.3
無*	8.0	12.8	20.5	13.8
有/無* (比)	3.08	2.87	2.47	2.70

無*: ROCCOシステムによる分類結果を利用

64

5.4.3 コーダによる評価

- 総合評価
- 主な機能評価
- 提示情報のわかりやすさ評価
- 操作性評価

65

総合評価

NANACOシステムの総合評価(単位: 人)

調査名	大変役に立った	やや役に立った	どちらともいえない	あまり役に立たなかった	全く役に立たなかった
JGSS-2003	7	4	0	0	0
2003年SSM調査	3	4	0	0	0
2005年SSM調査	17	3	0	0	0

参加コーダの属性 JGSS-2003: 学生(14人うち11人回答)

2003年SSM調査: 学生、教員(20人うち7人回答)

2005年SSM調査: 教員(35人うち20人回答)

66

主な機能評価

NANACO システムにおける主な機能別の評価（単位：人）

機能	大変役に 立った	やや役に 立った	どちらとも いえない	あまり役に 立たなかった	全く役に 立たなかった
職業コード候補の表示	12(2+10)	15(7+8)	4(2+2)	0	0
職業名の表示	20(5+15)	8(4+4)	2(2+0)	1(0+1)	0
職業内容の閲覧	14(5+9)	13(6+7)	3(0+3)	1(0+1)	0
職業定義内容の検索	23(9+14)	4(1+3)	3(1+2)	1(0+1)	0
データ検索	13(4+9)	9(1+8)	5(3+2)	2(1+1)	0

（ ）内の数字は「+」の前の数字がJGSS-2003、後ろが2005年SSM調査における人数

67

提示情報のわかりやすさ評価

表 6.6: 提示された情報のわかりやすさの評価（単位：人）

評価項目	大変わかり やすかった	ややわかり やすかった	どちらとも いえない	ややわかり にくかった	大変わかり にくかった
全体	8(2+6)	19(8+11)	4(1+3)	0	0

68

操作性評価

表 6.7: NANACO システムにおける操作性の評価（単位：人）

評価項目	大変わかり やすかった	ややわかり やすかった	どちらとも いえない	ややわかり にくかった	大変わかり にくかった
全体	10(5+5)	24(5+19)	1(1+0)	0	0
画面の切り替え	11(5+6)	9(2+7)	8(3+5)	3(1+2)	0
システム開始作業	7(1+6)	15(5+10)	6(4+2)	3(1+2)	0
システム終了作業	4(0+4)	9(5+4)	13(4+9)	4(2+2)	0

69

5.5 まとめ

- 提案した2つのアルゴリズムを統合し、コードの分類作業に役立つ情報システムとしてNANACOシステムを開発した
 - 社会調査における職業・産業コーディングの方法を大きく変え、その結果、調査票設計段階において職業データ収集の自由度が増え研究の推進に役立っている
- NANACOシステムは『SSM職業定義辞書』を変えることで、職業以外の自由回答にも適用が可能である
 - 例 SSM産業定義辞書
 - ISCO(国際標準職業分類)定義辞書
 - ISIC(International Standard Industrial Classification; 国際標準産業分類)定義辞書
 - ...
- 現在は、作業画面のレイアウトを利用する調査ごとに開発者側で作成しているが、今後は利用者側で対応してもらう予定である(HACHICOシステム)

70

6. おわりに

- 結論
 - 機械学習を用いて、自由回答を高精度に自動分類する方法および分類結果のクラス所属確率を高精度に推定する方法の提案を行った
 - また、上記のアルゴリズムを実現する情報システム(NANACOシステム)の開発を行い、職業コーディングにおいて利用が始まっている
- 今後の課題
 - [データ収集段階] 質のよいデータを収集するための支援方法を検討
 - 例えば、クラス所属確率を利用し、「効果的な」追加質問を行う
 - [コード決定段階] 人間によるチェックの必要度を判定できる方法を検討
 - [Domain Adaptation] 類似するドメインまたは類似するコード体系の事例に適合させる方法を検討

71

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

Kazuko TAKAHASHI, *PhD*¹ ; Hirofumi TAKI, *PhD*² ; Shunsuke TANABE, *PhD*³ ; Wei LI, *MD*⁴

¹Faculty of International Studies, KEIAI University, Japan. E-mail : takak@u-keiai.ac.jp

² Faculty of Social Science, HOSEI University, Japan

³ Faculty of Letters, Arts and Science, WASEDA University, Japan

⁴ Graduate School of Science and Engineering, TOKYO INSTITUTE OF TECHNOLOGY, Japan

Keywords : Answers to open-ended questions, Natural language processing, Machine learning

Introduction

The “**occupation and industry coding**” is a necessary task for statistical processing because respondent’s occupation and industry are collected as answers to open-ended questions in social surveys such as a national census. However, this task requires a great deal of labor and time-consuming. In addition, inconsistent results occur if the coders are not experts of coding. Our system assigns **three candidate codes** corresponding to the **National/International standard** to an answer by SVMs (Support Vector Machines), and attaches a **three-grade confidence level** to the first-ranked predicted code by using classification scores to support a manual check of the results. The system is now **open to the public** through the website of SSJDA (Social Science Japan Data Archive).

Occupation & Industry Code

● The National standard code used in Japanese social surveys

- **SSM occupation code** 200 classes
- **SSM industry code** 20 classes

*SSM : Social stratification and social mobility

● The International standard code defined by ILO

- **ISCO** (International standard classification of occupations) 400 classes
- **ISIC** (International standard industrial classification of all economic activities) 60 classes

Operation Process

Data File CSV format

ID, education, employment & job title, firm size, **job task, industry**

1001 9 2 8 to arrange the delivery vehicles load and upload of luggage

1002



Result File (e.g. SSM occupation code file) CSV format

ID	Confidence	Rank1	Rank2	Rank3	
1001	A	563	558	607	563 : a transportation clerk
1002					

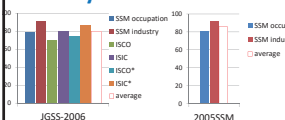
Experimental Setting

Training dataset : JGSS-2000,-2001,-2002,-2003,-2005 for SSM codes

2005SSM for ISCO & ISIC

Test dataset : JGSS-2006 (2005SSM only for SSM codes)

Accuracy of Each Kind of Codes (The Goal $\geq 80\%$)



*: in case using correct SSM codes instead of predicted codes

- The accuracy of only ISCO is under the value of the goal.
- The reasons are smallness of the dataset and large number of classes.
- A method using hierarchy of the ISCO structure may be effective in future (See the paper).

ACKNOWLEDGEMENTS

- JGSS (The Japanese General Social Survey) project
- The 2005 SSM Survey Research Group
- MEXT Grant-in-Aid for Scientific Research (C) 25380640

REFERENCES

- 1 Jung, Y., et al., 2008. A web-based automated system for industry and occupation coding. In *Proceedings of WISE 2008*, 443-457.

Related Work

- South Korea
A Web-based AIOCS (Automated System for Industry and Occupation Coding) ¹ with rule-based method, MEM and IR
- The United States
SOIC (Standardized Occupation & Industry Coding) ² with matching the rules according to the 1990 Census
NIOCCS (The NIOSH Industry & Occupation Computerized Coding System) ³ with matching the rules according to the 2000 Census
- Japan
ROCCO (Rule-based Occupation and Industry Coding) ⁴ for SSM codes
The combination method of SVMs and hand-crafted rules ⁵ for SSM codes

A Three-Grade Confidence Level

A (high) : (1) and (2), **B** (middle) : (1) and (3), **C** (low) : Otherwise

Score1 > 0 and Score2 <= 0 (1), Score1 - Score2 > Threshold (2)

Score1 - Score2 <= Threshold (3)

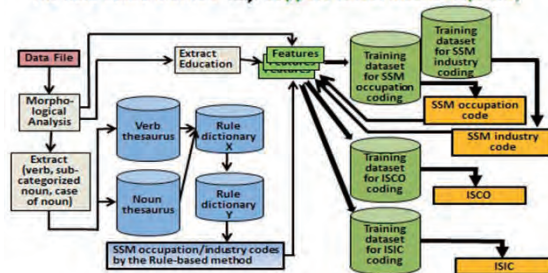
Score1 : the first-ranked score, Score2 : the second-ranked score

Open to the Public by SSJDA

<http://ssjda.iss.u-tokyo.ac.jp/joint/autocode/>

The Process of the System

The Rule-Based Method → Support Vector Machines (SVMs)



Effectiveness of a Three-Grade Confidence Level

- **Accuracy**
- **Coverage** (Threshold=3)
- Accuracy and Coverage in Level A are the most important criteria for coders.
- The accuracy of each kind of codes in Level A is higher than 94%. The accuracy of ISCO in level A is 96.3%, which is satisfactory.
- The coverage of each kind of codes in Level A is lower than 32%. The coverage of ISCO in level A is only 4.8%, which needs to be improved.

2 <http://www.cdc.gov/niosh/soic/default.html>

3 <http://www.cdc.gov/niosh-nioccs/>

4 Takahashi, K., 2000. A supporting system for coding of the answers from an open-ended question: An automatic coding system for SSM occupational data by case frame. *Sociological theory and methods* 15(1), 149-164.

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

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 -自然言語処理と機械学習を適用して-

高橋 和子 多喜 弘文 田辺 俊介 李 偉
敬愛大学 法政大学 早稲田大学 東工大

発表の順序

- ・ 研究の背景
- ・ 社会学における利用実績
- ・ 関連研究
- ・ 職業・産業コーディング自動化システム
- ・ Web公開版システムの利用方法
- ・ まとめ

研究の背景

- ・ 社会調査データには選択回答と自由回答あり
- ・ 自由回答により情報を得た場合、分類コードに変換する作業が必要
- ・ 選択回答が推奨されるが、**正確性**職業や産業情報は例外

◆ 地位	選択回答
◆ 役職	選択回答
◆ 従業先の規模	選択回答
◆ 仕事の内容(職業)	自由回答
◆ 従業先事業の内容(産業)	自由回答

地位・役職(選択回答)

JGSS-2003

- あなたの仕事は、大きく分けて、この中のどれにあたりますか。
- 1 経営者・役員
 - 2 常時雇用の一般従業者 役職なし
 - 3 " 職長、班長、組長
 - 4 " 係長、係長相当職
 - 5 " 課長、課長相当職
 - 6 " 部長、部長相当職
 - 7 " 役職はわからない
 - 8 臨時雇用・パート・アルバイト
 - 9 派遣社員
 - 10 自営業主・自由業者
 - 11 家族従業者
 - 12 内職
 - 13 わからない

従業先事業の内容(自由回答) JGSS-2003

- ▶ あなたが働いている場所(工場、事務所、商店、病院などの事業所)はどのような事業をしていますか。
例えば野菜の販売、自動車の製造、旅館、銀行の支店など、具体的にお聞かせください。

(できるだけ詳しく具体的に。会社名のみは不可。)

工 場

産業コーディング

産業コード

仕事の内容(自由回答) JGSS-2003

- ▶ あなたは通常、そこでどのような仕事をしていますか。仕事の内容を具体的にお聞かせください。(例えば、小学校教員、塾の講師、農作業、バスの運転、自動車の修理、スーパーのレジ、銀行の経理、コンピュータのプログラマー、営業事務、化粧品の外回り営業.....というように)

(できるだけ具体的に)

コピー機のトナーカートリッジの
製造

職業コーディング

職業コード

職業・産業コーディングの問題点と対策

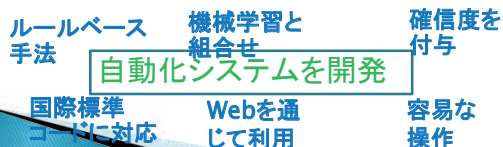
- ・分類クラスの多さとコーディングルールの複雑さ

→ コーダの作業負担大

職業コード約200 産業コード約20
自由回答以外の情報も用いた総合的な判断

- ・多人数による長期間の作業

→ コーディング結果における一貫性の問題



社会学における利用実績 (大規模調査)

- ▶ JGSS (Japanese General Social Surveys; 日本版総合的社会調査)
全米の総合的社会調査 (GSS) に範を取る二次分析のための生データ公開調査。
 - JGSS-2000, -2001, -2002, -2003, -2005, -2006, -2008, -2010 (EASS-2006, -2008, -2010)
 - 日本、台湾、韓国、中国
- ▶ SSM調査 (Social Stratification and social Mobility調査; 社会階層と社会移動全国調査)
社会階層や不平等、社会移動、職業、教育、社会意識などに関する社会調査。1955年以降10年ごとに実施。
 - 2005SSM調査, 2005SSM若年層調査, 2015SSM調査

社会学における利用実績 (大規模調査以外)

- ▶ 文部省科研費基礎研究A(2)「福祉社会の価値観に関する実証的研究」(研究代表 武川正吾(東大)) 2001年
- ▶ 東京大学社会科学研究所パネル調査「働き方とライフスタイルの変化に関する全国調査」2007年, 2008年, ..., 2013年
- ▶ 大阪大学人間科学研究科臨床死生学・老年行動学講座 権藤研究室調査2009年
- ▶ 成蹊大学アジア太平洋研究センター「暮らしについての西東京市民アンケート」(代表 小林盾) 2009年, 2010年, 2011年, 2012年
- ▶ 平成22年度 二十一世紀文化学術財団奨励金「結婚と子育て支援に関する東京都民調査」(代表 金井雅之(専修大)) 2012年
- ▶ 平成22年度～平成24年度 文部省科研費基盤研究(B)「地域間格差と個人間格差の調査研究: ソーシャルキャピタル論的アプローチ」(代表 辻竜平(信州大)) 2012年
- ▶ 等々

- ▶ Web公開版の利用は今年度12件

関連研究(職業・産業コーディング)

- ▶ 韓国 大韓民国統計庁
Web-based AIOCS (Y. Jung, J. Yoo, S-H. Myaeng and D-C. Han, 2008)
 - ・ ルールベース手法→最大エントロピー法→情報検索技術 正解率76.3%
 - ・ 1問1答方式(会社名、ビジネスカテゴリ、部門、役職、仕事の内容)
- ▶ 米国 CDC (Centers for Disease Control and Prevention)
アメリカ疾病予防管理センター
SOIC (Standard Occupation & Industry Coding)
<http://www.cdc.gov/niosh/soic/SOIC.About.html>
 - ・ 単語のマッチングが主
 - ・ 正解率 職業75% 産業76% 職業&産業63%
 - ・ ソフトウェアをダウンロードして利用
- NIOSCS**
(The NIOSH Industry & Occupation Computerized Coding System)
<http://www.cdc.gov/niosh-nioscs/>
 - ・ ルールベース手法
 - ・ 1問1答方式 または ファイルによるデータの受け渡し
 - ・ 結果に3段階の確信度付与 (High, Medium, Low)

機械学習を適用せず

SSJDAによる公開版

職業・産業コーディング自動化システム

- ▶ 国内・国際標準の職業・産業コード計4種類に変換
- ▶ 入力: 職業・産業情報をもつ所定の形式のファイル (CSV形式)
- ▶ ルールベース手法を機械学習 (SVM) に組み込んだ手法
- ▶ 出力: コードの種類ごとにSVMにより予測された結果のファイル (CSV形式)
- ▶ 3段階の確信度付与
A: コーダの作業不要 B: できればコードの作業必要
C: コーダの作業必要
- ▶ Webを通じてだれでも利用可能
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) より試行提供中
- ▶ だれでも容易に操作可能

4種類のコード

	コードの種類	コードの数	備考
国内標準	SSM職業コード (小分類)	約200	501～688 700番台、800番台も追加
	SSM産業コード (大分類)	約20	10、20、91、92、93、...、170 81、82、171、172も追加
国際標準	ISCO (小分類)	約400	4桁 (大分類、亜大分類、中分類、小分類)
	ISIC (亜大分類)	約60	4桁 (大分類、亜大分類、中分類、小分類)

ILO

International Standard Classification of Occupations

International Standard Industrial Classification of All Economic Activities

入力データファイル例

ID	学歴	地位・役職	従業先事業の内容	仕事の内容	事業規模
1	9	9	工場	コピー機のトナーカートリッジの製造	8
2	9	3	工場	ガラス吹き	6
3	11	4	福祉事務所	生活保護業務の現業員	9
4	11	8	予備校	事務	8
5	10	2	病院	看護師	4

結果ファイル例 (SSM職業コードの場合)

ID	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

630: 金属工作機械工、めっき工、金属加工作業者
631: 鉄工、板金工

確信度

自動コーディングの結果(第1位)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したもの

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

複数のスコアを用いる

A: 第1位のスコア>0 第2位のスコア<=0
第1位のスコア-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-第2位のスコア<= α

C: A、B以外

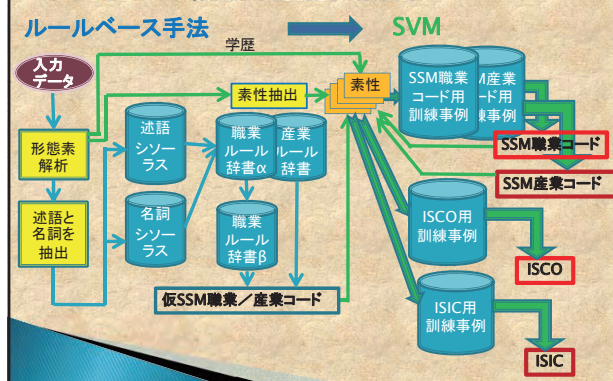
α は閾値

(今回 $\alpha=3$)

自動化の手法

	コードの種類	自動化の手法(SVMで用いる素性)
国内標準	SSM職業コード(小分類)	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)	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)
国際標準	ISCO(小分類)	SVM(基本素性, SVMにより第1位に予測されたSSM職業コード, 学歴)
	ISIC(亜大分類)	SVM(基本素性, SVMにより第1位に予測されたSSM産業コード)

システムの構成(処理の流れ)



ルールベース手法

格フレームの概念を利用した情報抽出

必要な情報だけ利用

- 述語を抽出
- 分類に必要な表層格に該当する語(名詞)を抽出
- 三つ組みを生成

例 仕事の内容「大学で哲学を教えている」

大学で哲学を教えている

(大学で教える)

名詞 表層格 述語

ルールベース手法

シソーラスによる語の拡張

三つ組み中の述語と名詞を拡張

▶ 述語シソーラス

- ・教える 教え込む ... 述語コード「364 1」
- ・作る 製造 製作 ... 述語コード「386 1」

見出し語 10,871語

▶ 名詞シソーラス

- ・(電気機械等 電気機器 カデンピン 家電 空調 クーラー エアコン テレビ 液晶テレビ TV 冷蔵庫 洗濯機 ...)

見出し語 330語

ルールベース手法

ルール辞書によるコードの決定

三つ組みからコードを決定するためのルール

- ▶ 職業ルール辞書(産業ルール辞書も同様)
((述語コード)(SSM職業コード(表層格 名詞))

((386 1)(506 (を ソフトウェア システム))
...
(599 (を 作物 野菜 果樹 蚕種))
...
(625 (を ガラス製品 セメント製品))
(626 (を その他窯業 ガラスウール))
...
(704 ()) 78個の職業コード

* SSM職業コードによっては、この後「地位・役職」「事業規模」もチェックして最終的に決定

システムの評価

▶ 開発者による評価 4種類のコード

- ・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
- ・確信度の有効性

▶ 利用者による評価 SSM職業コード

- ・「自動コーディングシステムをそのまま用いた場合の結果」と、「正解」との間のズレを検討することで、本システムをそのまま適用した場合の有効性と限界を明らかにする
- ・その試みを通じて、SSM職業分類の特徴やコードに必要なスキルについても考察する

正解...最終的に人手で付与されたコード

開発者による評価(実験設定)

4種類のコード別に評価

- ▶ 国内標準コード 39,120事例

JGSSデータセット(2000年~2003年)	SSM調査データセット(2005年)
JGSSデータセット(2005年)	16,089事例
JGSSデータセット(2006年)	
2,203事例	

- ▶ 国際標準コード

JGSSデータセット(2006年)	SSM調査データセット(2005年)
2,203事例	16,089事例

正解率(第3位まで)

コード	JGSS-2006	2005SSM
SSM職業コード	78.8%	80.6%
SSM産業コード	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%	—

確信度別正解率とカバー率

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%)

利用者による評価(実験設定)

訓練事例

- JGSS-2000, -2001, -2002, -2003, -2005, -2006, -2008, -2010(33,712事例)+2005SSM(16,083事例) 計49,795事例

評価事例

- 東京大学社会科学研究所が実施する「働き方とライフスタイルの変化に関する全国調査」(若年・壮年パネル調査:JLPS)の第1波 3,619事例

実際の研究で利用される状況で評価(SSM職業コード)

- ・確信度の有効性
- ・SSM職業大分類
- ・SSM新総合8分類
- ・職業威信スコア(職業の経済的地位と社会的地位を総合した客観的な地位の推定値)

実際の研究で使われる分類

SSM職業大分類

- 専門 管理 事務 販売 熟練 半熟練 非熟練 農林

SSM総合職業分類

- 専門 管理 大企業ホワイト 中小企業ホワイト 大企業ブルー 中小企業ブルー 自営ノンマニュアル 自営マニュアル 農業

SSM新総合8分類

- 専門 大企業ホワイト 中小企業ホワイト 自営ホワイト 大企業ブルー 中小企業ブルー 自営ブルー 農業

自動コーディングの正解率 (SSM職業大分類(旧職業8分類))

小分類	職業大分類	改善率	(割合)	確信度	小分類	職業大分類 (旧職業8分類)
専門	77.7%	85.2%	7.5% (23.0%)			
管理	57.1%	57.1%	0.0% (0.8%)			
事務	59.5%	75.7%	16.2% (28.2%)	A	97.8	98.6
販売	72.2%	80.7%	8.6% (18.0%)	B	76.4	84.7
熟練	66.9%	75.4%	8.6% (15.2%)	C	40.2	56.7
半熟練	63.4%	72.6%	9.2% (9.4%)	全体	66.9	77.5
非熟練	49.4%	57.1%	7.7% (4.5%)			
農林	81.3%	81.3%	0.0% (0.9%)			
合計	67.2%	77.5%	10.3% (100.0%)			

- ▶ 大分類にすることで、約10%正解率が上がった
- ▶ 特に事務の正解率は大幅に上がっている
- ▶ 非熟練の正解率はあまり高くないまま

自動コーディングの正解率 (SSM総合職業分類(簡略版9分類))

小分類	総合職業分類 簡略版	改善率	(割合)	確信度	小分類	総合職業 分類 簡略版
専門	77.7%	86.6%	8.9% (23.1%)			
管理	60.0%	63.3%	3.3% (0.9%)	A	97.8	99.4
大企業ホワイト	68.1%	87.9%	19.8% (12.5%)	B	76.4	92.2
中小企業ホワイト	63.3%	86.4%	23.0% (32.0%)	C	40.2	71.1
大企業ブルー	62.0%	82.5%	20.5% (3.6%)	全体	66.9	86.0
中小企業ブルー	63.2%	85.7%	22.5% (22.9%)			
自営ノンマニュアル	58.7%	82.5%	23.8% (1.9%)			
自営マニュアル	62.6%	86.9%	24.3% (3.2%)			
農業	81.3%	81.3%	0.0% (0.9%)			
合計	67.0%	86.0%	19.0% (100.0%)			

- ▶ 職業大分類よりも9%くらい正解率が上昇
- ▶ 確信度Bでも9割を超える正解率に
- ▶ ブルーカラーのスキルを区別しなくなったことが大きい

自動コーディングの正解率 (新総合 8分類)

小分類	職業大分類	改善率	(割合)	確信度	小分類	新総合 8分類
専門	77.7%	86.6%	8.9% (22.9%)			
大企業ホワイト	65.3%	88.4%	23.0% (18.7%)	A	97.8	99.6
中小企業ホワイト	64.1%	87.5%	23.4% (25.3%)	B	76.4	92.7
自営ホワイト	59.1%	86.0%	26.9% (2.7%)	C	40.2	72.3
大企業ブルー	62.0%	83.6%	21.6% (6.5%)	全体	66.9	86.7
中小企業ブルー	63.5%	86.3%	22.8% (19.1%)			
自営ブルー	62.1%	84.1%	22.0% (3.9%)			
農業	81.3%	81.3%	0.0% (0.9%)			
合計	67.2%	86.8%	19.6% (100.0%)			

- ▶ これまで最も正解率が高い(管理がなくなった分向上)
- ▶ ボリュームの大きいカテゴリの正解率がいずれも85%を超えている

職業威信スコアの平均値

職業威信スコア

職業の経済的地位と社会的地位を総合した客観的な地位の推定値

- ▶ 年齢別、学歴別に平均値を求めても、正解と自動コードの威信スコアの平均値にほとんど違いはみられない(すべてのカテゴリで95%信頼区間が重なる)

	正解	自動コード		正解	自動コード
20～25歳	48.86	49.20	中学校	45.5	46.1
26～30歳	51.94	52.06	高等学校	47.4	48.0
31～35歳	52.16	52.01	専門学校	51.9	51.9
36～40歳	51.84	52.05	短大・高専	51.6	51.7
全体	51.24	51.36	大学	52.7	52.7
			大学院	63.8	62.4
			全体	51.2	51.4

職業威信スコアの相関

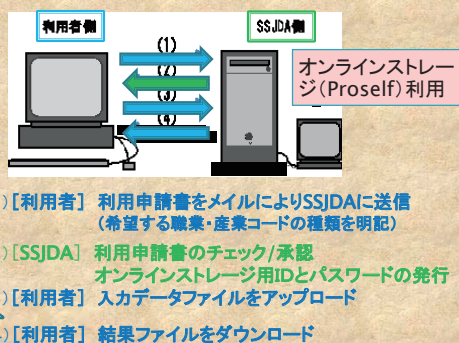
	正解威信	自動コード威信
父職威信スコア	0.125	0.093
(95%信頼区間)	(0.088~0.162)	(0.056~0.130)
本人教育年数	0.350	0.296
(95%信頼区間)	(0.319~0.381)	(0.264~0.328)

- ▶ 正解と自動コード威信スコアの相関係数は0.788
- ▶ 正解の方が、父職の威信スコアや本人の教育年数との相関が高い
- ▶ ただし、これについても標準誤差から信頼区間を求めると、重なりがあるため数値が異なっているとはいえない

利用者による評価(まとめ)

- ▶ 確信度Aのものは小分類、大分類ともにほぼそのまま使える
- ▶ 「専門職」の正解率は高いが、「管理職」「事務職」「生産現場・技能職」の小分類正解率は6割以下とやや低い
- ▶ このうち、事務系の職業は小分類ではなく「事務職」というカテゴリでくると大幅に正解率が向上する → 職務が明確でない日本の雇用の特徴か
- ▶ ブルーカラー職はスキルレベルで分けた場合(職業大分類)正解率が低い
- ▶ 総合職業分類や新総合8分類は、企業規模と雇用形態を重視するので自動コードには有利。確信度Bでも9割を超える
- ▶ 職業威信も少なくとも平均値を比較するような場合はほとんど問題なさそう

Web公開版システムの利用方法



SSJDAのWebサイト



自動コーディングの説明



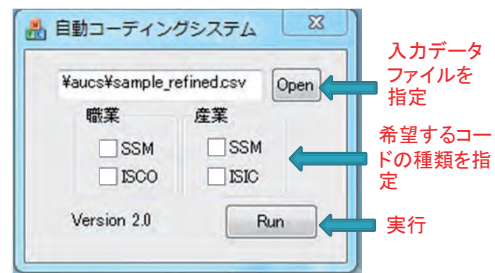
利用方法の説明



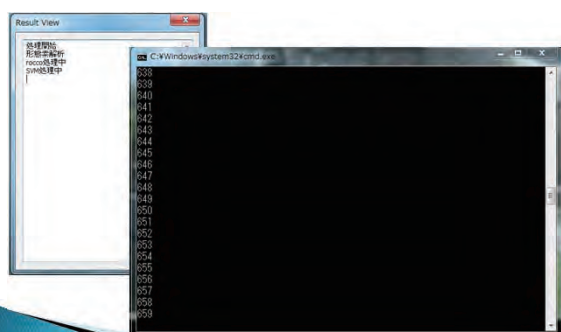
入力ファイルの説明



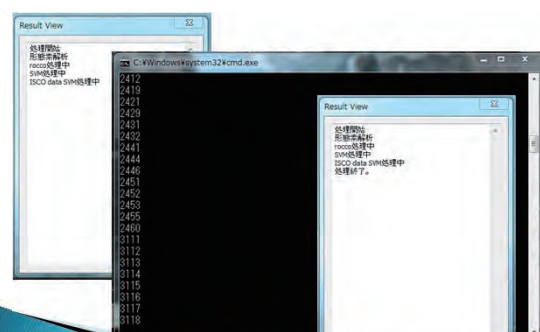
SSJDA側 システム操作画面



処理過程の画面 (SSM職業コード処理)



処理過程の画面 (ISCO処理)



まとめ

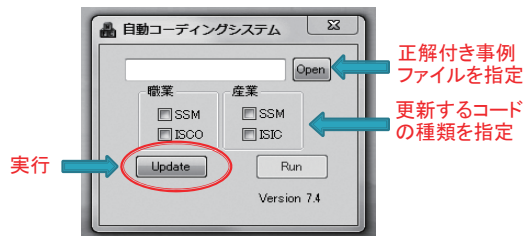
- 自由回答のコーディング支援例として、社会学で利用されている「機械学習を適用した職業・産業コーディング自動化システム」について報告
 - 国内／国際標準の4種類のコードに対応
 - 職業は約7割から8割、産業は約8割から9割の正解率
 - 予測第1位の候補に付与される確信度の有効性を確認
 - 東大社会科学研究所SSJDAのWebを通じて利用
 - システムの操作は容易
- メンテナンスの自動化機能として、**訓練事例追加機能**を追加中

メンテナンス自動化機能(入力ファイル)

- 正解付き事例の蓄積 → **訓練事例の追加**
システムの永続性の点から容易な操作

ID	学歴	地位・役職	従業先事業の内容	仕事の内容	事業規模	SSM職業コード	SSM産業コード	ISCO	ISIC
1	9	9	工場	コピー機のトナーカートリッジの製造	8	630	60		
2	9	3	工場	ガラス吹き	6	625	60		

メンテナンス自動化機能(操作画面)



まとめ

- ▶ 自由回答のコーディング支援例として、社会学で利用されている「機械学習を適用した職業・産業コーディング自動化システム」について報告
 - 国内／国際標準の4種類のコードに対応
 - 職業は約7割から8割、産業は約8割から9割の正解率
 - 予測第1位の候補に付与される確信度の有効性を確認
 - 東大社会科学研究所SSJDAのWebを通じて利用
 - システムの操作は容易
- ▶ メンテナンスの自動化機能として、訓練事例追加機能を追加中
- ▶ 自由回答一般への拡張方法を検討中

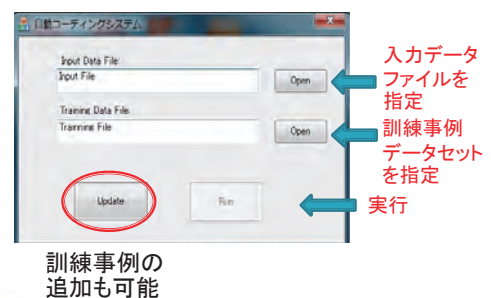
システムの拡張

- ▶ 訓練事例を準備すれば本システムの拡張可能
- ▶ SVMを適用
- ▶ 入力データファイルの形式

ID	質問1	質問2	質問3	質問4	質問5
番号	選択回答	選択回答	選択回答	自由回答	自由回答

最大選択回答3種類、自由回答2種類までの情報
 訓練事例の生成は**訓練事例の追加機能**を利用

システムの拡張(操作画面)



訓練事例の追加も可能

まとめ

- ▶ 自由回答のコーディング支援例として、社会学で利用されている「機械学習を適用した職業・産業コーディング自動化システム」について報告
 - 国内／国際標準の4種類のコードに対応
 - 職業は約7割から8割、産業は約8割から9割の正解率
 - 予測第1位の候補に付与される確信度の有効性を確認
 - 東大社会科学研究所SSJDAのWebを通じて利用
 - システムの操作は容易
- ▶ メンテナンスの自動化機能として、訓練事例追加機能を追加中
- ▶ 自由回答一般への拡張方法を検討中
- ▶ 質のよい入力データ収集方法
 - 調査現場で利用する **入力支援システム?**

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。
- ▶ 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 25380640の助成を受けたものである。

システム開発に関連するこれまでの研究課題・研究組織と研究概要

(所属等は異動も含め研究期間当時のもの)

(1)

研究種目 一般研究 (C)

研究課題題名 SSM 職業コーディング支援エキスパートシステムの構築

研究番号 07610221

研究期間 平成 7 (1995) 年度

研究代表者

高橋和子 千葉敬愛短期大学国際教養科・助教授

研究概要

従来、人手で行ってきた SSM 職業コーディングをコンピュータで支援させるためには、用語が担う「意味内容」を可能な限りコンピュータの理解できる「形式」に置き換える必要がある。今年度は実際の SSM 職業データを用いて、回答の内容を分析した。以下、具体的に述べる。

1. 1995 年 SSM 調査委員会の好意により入手可能となった 1965 年データ (東京地区) について、全サンプルの入力を完了した。2. 上記データのうち、先に入力していた 1/4 については、回答の形態や出現する用語の傾向を分析した。3. この結果を、カテゴリーの説明書である『SSM 職業分類 (改訂版)』における記載内容と比較し、対応辞書の作成に向けて両者の関連を部分的に明らかにした。今後は、全サンプルについて同様の分析を行って基本的な戦略を立てる。さらに可能ならば、最新版 (1995 年) データの利用許可を得て分析し、精度を上げていきたい。

(2)

研究種目 基盤研究 (C)

研究課題題名 自然言語処理技術を適用した自由回答コーディング支援システムの開発

研究番号 16530341

研究期間 平成 16 (2004) 年度～平成 17 (2005) 年度

研究代表者

高橋和子 敬愛大学国際学部国際学科・助教授

研究分担者

高村大也 東京工業大学精密工学研究所・助手

研究概要

研究成果は、社会調査において代表的な自由回答である職業データのコーディング (「職業コーディング」) において、コーダ (人間) を支援する「NANACO システム」を完成させたことである。システムは職業データに限定されるものではなく、「分類カテゴリーをもつ」自由回答であれば適用可能である。1. まず、支援するための最優先課題は、事例に対して高精度なコー

ディング（分類）結果を提示することである。このために、自然言語処理分野の中でも特に「文書分類」における先端的な技術を応用し、職業分類コードを決定するルールを辞書にまとめて利用する手法（「ルールベース手法」）と、過去の事例を訓練データとして用いる「機械学習による手法」を有機的に組み合わせた手法を考案し有効性を示した。2. 次に、システムを利用したコード達からもっとも要請の高かった「分類結果に対するシステムの確信度（クラス所属確率）」を付与した。クラス所属確率の推定をできる限り正確に行う方法として、訓練データにおける複数次元の分類スコアを用いて作成した「正解率表」を利用する方法を考案し有効性を示した。3. さらに、コードの作業そのものを支援するために、「分類カテゴリーの定義ファイルの閲覧」「コーディング時に参照したいデータの表示」「注意マークの付与」などの機能も充実させた。4. 現在は提示情報のすべてをバッチにより作成しているが、今後は一部を Web 処理に移行する予定である。5. システムは、2003 年 SSM 予備調査、JGSS-2005 調査、2005 年 SSM 調査（産業コーディングも）に適用された。今後、2005 年 SSM 若年層調査（2007 年 3 月）、JGSS-2006 調査（ISCO によるコーディングを含む）（2007 年 5 月）が予定されている。

（3）

研究種目 基盤研究（C）

研究課題題名 社会調査の基盤を提供する自由回答の自動コーディングシステムの開発と公開

研究番号 22530516

研究期間 平成 22（2010）年度～平成 24（2012）年度

研究代表者

高橋和子 敬愛大学国際学部国際学科・教授

研究分担者

田辺俊介 東京大学社会科学研究所・准教授

吉田 崇 東京大学社会科学研究所・助教 → 静岡大学人文社会科学部・准教授

研究協力者

魏 大比 東京工業大学大学院情報理工学研究科・研究員 → 名校教育グループ・代表

李 偉 東京工業大学大学院理工学研究科・博士課程在学

研究概要

社会調査では回答者の職業や産業は重要で、正確さを期するために自由回答で収集するケースが多い。しかし、統計処理のために収集後にコード化する作業が必須で、最近では国内標準コードに加えて国際標準コードの要請も生じており、コードの負担が増大している。本研究では、自然言語処理や機械学習など人工知能における最新の成果を適用してコーディング作業を自動化し、結果を Web により入手できるシステムを開発した。その際、各コードには人間による見直しが必要か否かを 3 段階の確信度で付与するため、作業の大幅な軽減が見込める。

システム開発に関連するこれまでの主要な研究成果（テーマ別）

ルールベース手法

- 高橋和子, 1998, 「自然言語処理による SSM 職業コーディングの自動化システム」 盛山和夫（編）『現代日本の社会階層に関する全国調査研究成果報告書 SSM 調査シリーズ [5] 職業評価の構造と職業威信スコア』（科学研究費補助金特別推進研究成果報告書）, 1995 年 SSM 調査研究会, pp.195-228.

http://srdq.hus.osaka-u.ac.jp/PDF/SMM1995_r5_11.pdf

- 高橋和子, 2000, 「自由回答のコーディング支援について一格フレームによる 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.
- 高橋和子, 2002, 「JGSS-2000 における職業・産業コーディング自動化システムの適用」『日本版 General Social Surveys 研究論文集 JGSS-2000 で見た日本人の意識と行動 [東京大学社会科学研究所資料第 20 集]』, pp.171-184.

http://jgss.daishodai.ac.jp/research/monographs/jgssm1/jgssm1_13.pdf

- 高橋和子, 2003, 「JGSS-2001 における職業・産業コーディング自動化システムの適用」『日本版 General Social Surveys 研究論文集 [2] JGSS で見た日本人の意識と行動 [東京大学社会科学研究所資料第 21 集]』, pp.179-192.

http://jgss.daishodai.ac.jp/research/monographs/jgssm2/jgssm2_11.pdf

機械学習

- 高橋和子・高村大也・奥村学, 2004, 「ルールベース手法と機械学習による自由回答の分類ー職業コーディング自動化の方法ー」『理論と方法』Vol.19, No.2, pp.177-196.

https://www.jstage.jst.go.jp/article/ojjams/19/2/19_2_177/_pdf

ルールベース手法と機械学習の組合せ手法

- 高橋和子, 2004, 「職業コーディングにおける ROCCO システムと SVM の組み合わせ」『日本版 General Social Surveys 研究論文集 [3] JGSS で見た日本人の意識と行動 [東京大学社会科学研究所資料第 24 集]』, pp.163-174.

http://jgss.daishodai.ac.jp/research/monographs/jgssm3/jgssm3_12.pdf

- Kazuko TAKAHASHI and Hiroya TAKAMURA and Manabu OKUMURA, 2005, “Automatic Occupation Coding with Combination of Machine Learning and Hand-Crafted Rules.” *The 9th International Conference on Pacific-Asia Knowledge Discovery and Data Mining 2005 (PAKDD'05). Lecture Notes in Artificial Intelligence, Vol.3518*, pp.269-279, Springer-Verlag.

http://link.springer.com/chapter/10.1007%2F11430919_34#page-1 (一部のみ掲載)

- 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.

https://www.jstage.jst.go.jp/article/jnlp1994/12/2/12_2_3/_pdf

- 高橋和子, 2007, 「産業・職業分類自動コーディングの開発と活用」独立行政法人統計センター研究センター平成 19 年度第 1 回データエディティング研究会講演会講演 (2007 年 12 月 12 日) .
- 高橋和子, 2008, 「29 章 コーディングの自動化」谷岡一郎・仁田道夫・岩井紀子 (編著)『日本人の意識と行動 日本版総合的社会調査 JGSS による分析』東大出版会, pp.459-471.

ISCO 処理

- 高橋和子, 2008, 「機械学習による ISCO 自動コーディング」前田忠彦 (編)『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』(科学研究費補助金特別推進研究成果報告書), 2005 年 SSM 調査研究会, pp.47-68.

(再掲 2013 年, 佐藤嘉倫 (監修)『社会階層調査研究資料集—2005 年 SSM 調査報告書—第 1 巻 SSM の基礎分析と社会調査の諸問題』日本図書センター)

- 高橋和子, 2011, 「ISCO 自動コーディングシステムの分類精度向上に向けて—SSM および JGSS データセットによる実験の結果—」大阪商業大学 JGSS 研究センター編『JGSS Research Series No.8: 日本版総合的社会調査共同研究拠点研究論文集[11]』, pp.193-205.

http://jgss.daishodai.ac.jp/research/monographs/jgssm11/jgssm11_17.pdf

確信度の決定 (クラス所属確率の推定)

- Kazuko TAKAHASHI and Hiroya TAKAMURA and Manabu OKUMURA, 2007, “Estimation of Class Membership Probabilities in the Document Classification.” *The 11th International Conference on Pacific-Asia Knowledge Discovery and Data Mining 2007 (PAKDD’07). Lecture Notes in Artificial Intelligence, Vol.4426*, pp.284-295, Springer-Verlag.

http://link.springer.com/chapter/10.1007%2F978-3-540-71701-0_29#page-1 (一部のみ)

- Kazuko TAKAHASHI and Hiroya TAKAMURA and Manabu OKUMURA, 2008, “Direct estimation of class membership probabilities for multiclass classification using multiple scores.” *Knowledge and Information Systems, 19(2)*, pp.185-210, Springer London.

<http://link.springer.com/article/10.1007/s10115-008-0165-z#page-1> (一部のみ)

- 高橋和子・高村大也・奥村学, 2008, 「複数の分類スコアを用いたクラス所属確率の推定」『自然言語処理』Vol.15 No.2, pp.3-38.

https://www.jstage.jst.go.jp/article/jnlp1994/15/2/15_2_3/_pdf

- 高橋和子・田辺俊介・吉田崇・魏大比・李偉, 2013, 「確信度付き職業・産業コーディング自動化システムの開発と公開」『数理社会学会第 55 回大会報告要旨集』, pp.38-41. (2013 年 3 月 19 日 於: 東北学院大学)

Web による公開と容易な操作

- 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.
http://www.anlp.jp/proceedings/annual_meeting/2012/pdf_dir/P1-7.pdf
- 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.
http://www.anlp.jp/proceedings/annual_meeting/2013/pdf_dir/P5-8.pdf

アンサンブル学習による精度の向上

- 高橋和子, 2009, 「クラス所属確率を用いた事例ごとの分類器選択」『言語処理学会第 15 回年次大会発表論文集』, pp.709-712. (2009 年 3 月 11 日 於: 鳥取大学)
http://www.anlp.jp/proceedings/annual_meeting/2009/pdf_dir/D4-8.pdf
- 高橋和子, 2009, 「サポートベクターマシンにおけるアンサンブル学習の提案」『人工知能学会第 23 回大会発表論文集』. (2009 年 6 月 19 日 於: 高松サンポートホール)
<https://kaigi.org/jsai/webprogram/2009/pdf/102.pdf>
- 高橋和子, 2010, 「クラス所属確率を利用したアンサンブル学習」『人工知能学会第 24 回大会発表論文集』. (2010 年 6 月 9 日 於: 長崎ブリックホール)
<https://kaigi.org/jsai/webprogram/2010/pdf/260.pdf>
- 高橋和子, 2011, 「クラス所属確率を用いた多クラス SVM におけるアンサンブル学習」『情報処理学会第 73 回全国大会論文集』, pp.2-25—2-26. (2011 年 3 月 4 日 於: 東工大大岡山キャンパス)

NANACO システム

- 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 <JGSS Research Series No.1>』, pp.225-242.
http://jgss.daishodai.ac.jp/research/monographs/jgssm4/jgssm4_13.pdf

あとがき

今回の科学研究費補助金により、「職業・産業コーディング自動化システム」を一応、完成させることができた。

振り返ると、1995 年 SSM 調査のコーディング合宿に参加し、この膨大な作業をコンピュータで何とか支援できないものかと考えてから 20 年の歳月が流れたことになる。この間、本システムは多くの方に支えられながら、少しずつ成長してきた。

初めは原純輔先生がお持ちの知識をすべてコンピュータに入れたシステムを作ろうと思い、東北大学まで『社会調査演習』を手にとったところ、掲載されている職業・産業コーディングの例を一つ一つ丁寧に解説して下さったこと、また、実データによる実験のため、本郷キャンパスの盛山和夫先生のところに通い、ダンボール箱から調査票を取り出しては入力させていただいたことも今ではなつかしい思い出である。そして、まだ実験段階にあったシステムを、石田浩先生（東大社会科学研究所）が利用してみたいとおっしゃったことは、実用化に向けた大きな一歩となり、励みとなった。

その後、精度向上のために機械学習を取り入れたアルゴリズムの開発では、奥村学先生と高村大也先生のご指導を始め、東工大奥村研究室の方達に大変お世話になった。このような機会を得ることができたのは、サバティカルとして 1 年間、研究に専念できる環境を与えてくれた勤務校（敬愛大学）のおかげである。

また本システムの開発については、実際にシステムの利用先として想定された日本版総合社会調査（JGSS）と社会階層と社会移動に関する全国調査（SSM）に対しても、その関係者の皆様のみならず、その調査の存在自体に、ここに謝意を示しておきたい。

本システムはまだまだ検討すべき点が残されているが、今後は、自動コーディングの根幹部分についての大幅な変更は行わず、利用者にとって便利な機能の追加や改良を行っていく予定である。

また、コンピュータだけでなく人間にとっても、コーディングの精度を高めるには、その前段階である調査現場で収集されるデータの質を向上させる必要性を強く感じており、このための支援システムの開発も計画している。

本システムは、人間が機械に学習すべき素性を与え、学習のための正解付き事例を用意する教師付き学習であるが、将来は、昨今、話題となっている深層学習（Deep Learning）が職業・産業コーディングにも適用される日が来るのかもしれない。

最後に、今回およびこれまでの共同研究者や研究協力者の方々はもとより、システムの試行提供などの面で大変お世話になった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社会調査分野の藤原翔先生と石田賢示先生に深く感謝いたします。

研究課題・研究組織

(所属等は異動も含め研究期間のもの)

研究種目 基盤研究 (C)

研究課題題名 社会調査の基盤を提供する自動コーディングシステムの Web 提供 :
その国際化と汎用化

研究番号 25380640

研究期間 平成 25 (2013) 年度～平成 27 (2015) 年度

研究代表者

高橋和子 敬愛大学国際学部国際学科・教授

〒263-8588 千葉県稲毛区穴川 1-5-21

e-mail : takak@u-keiai.ac.jp

研究分担者

田辺俊介 早稲田大学文学学術院・准教授

多喜弘文 東京大学社会科学研究所・助教 → 法政大学社会学部・講師

研究協力者

李 偉 東京工業大学大学院理工学研究科・博士課程在学

研究成果一覧

国際会議論文

- Kazuko TAKAHASHI and Hirofumi TAKI and Shunsuke TANABE and Wei LI, 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”, *Proceedings of the 6th International Conference on Knowledge Engineering and Ontology Development (KEOD 2014)*, DOI: 10.5220/0005131703690375, pp.369-375.
https://www.researchgate.net/publication/276090302_An_Automatic_Coding_System_with_a_Three-Grade_Confidence_Level_Corresponding_to_the_NationalInternational_Occupation_and_Industry_Standard_Open_to_the_Public_on_the_Web (一部のみ)

学会発表論文

- 高橋和子・多喜弘文・田辺俊介・李偉, 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』, pp.932-935.
(2014 年 3 月 20 日 於：北海道大学) .
http://www.anlp.jp/proceedings/annual_meeting/2014/pdf_dir/P8-15.pdf
- 高橋和子・多喜弘文・田辺俊介・李偉, 2016, 「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—」『情報処理学会第 78 回（平成 28 年）全国大会講演論文集（4）』, pp. 495-496. (2016 年 3 月 10 日 於：慶應大学) .
- 高橋和子・多喜弘文・田辺俊介・李偉, 2016, 「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—自然言語処理と機械学習の適用—」言語処理学会第 22 回年次大会ワークショップ「言語処理の応用」. (2016 年 3 月 11 日 於：仙台国際センター) .
http://www.anlp.jp/proceedings/annual_meeting/2016/workshop1/pdf/ws1.pdf

学会発表

- 高橋和子・多喜弘文・田辺俊介・李偉, 2014, 「職業・産業コーディング自動化システムの一般公開に向けた課題と対応」『数理社会学会第 57 回大会報告要旨集』, pp.68-71. (2014 年 3 月 8 日 於：山形大学) .
- 高橋和子, 2015, 「職業コーディング自動化システム評価（得意／苦手な分類）」『数理社会学会第 59 回大会報告要旨集』, pp.76-79. (2015 年 3 月 14 日 於：久留米大学) .
- 高橋和子・多喜弘文・田辺俊介, 2016, 「職業・産業コーディング自動化システムの利用に関する評価—社会階層研究を事例に」『数理社会学会第 61 回大会報告要旨集』, pp.31-34.
(2016 年 3 月 17 日 於：上智大学) .

2016 年 3 月

平成 25～27 年度科学研究費補助金基盤研究 (C) 25380640

職業・産業コーディング自動化システム

An Automatic Occupation and Industry Coding System

高橋和子 (敬愛大学国際学部教授)

電話 : 043-251-6363 (代) FAX : 043-251-6407

E-mail : takak@u-keiai.ac.jp