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―

高橋 和子
敬愛大学国際学部

Improvement of Classification Accuracy in an ISCO Automatic Coding System:
Results of Experiments Using both the SSM Dataset and the JGSS Dataset

Kazuko TAKAHASHI
Faculty of International Studies
Keiai University

In social surveys, we need to conduct the occupation coding when occupation data is obtained by open-ended questionnaire. Conducting the occupation coding manually is a time-consuming and complicated task and sometimes leads to inconsistent coding results when coders are not experts.

For this reason, the automatic coding system, which is a combination of a rule-based method and Support Vector Machines (SVMs), has been developed and used for SSM occupation codes, which are usually used in Japanese social surveys. Recently, coders are often requested to conduct both the SSM occupation coding and the ISCO (International Standard Classification of Occupation) coding. Therefore an automatic coding system for ISCO codes should be also developed. The purpose of this paper is to report results of experiments designed for improvement of classification accuracy in the ISCO automatic coding system by using real datasets from the 2005SSM surveys and the JGSS.

Key Words: JGSS, ISCO automatic coding system, Support Vector Machines

社会調査において重要な「職業」の情報が自由回答の場合は、統計処理を可能にするためのコード化が必要である。国際比較研究の活発化に伴い、最近は国内標準の SSM 職業分類コードだけでなく国際標準職業分類（ISCO コード）の付与も要請されているが、2 種類のコーディングはコーダの負担や作業時間をこれまで以上に増大させる。一つの解決策は、ISCO コーディングに対しても、既開発の SSM 職業自動コーディングシステムと同様にコンピュータによる自動化を行ってコーダを支援することである。ISCO コーディングの場合は SSM 職業コーディングと状況が異なるために、全く同じアルゴリズムは適用できず、機械学習に限定した方法を検討している。今回、現実のデータである SSM および JGSS のデータセットを用いた実験により、素性選択や訓練データ作成に関して現時点での自動化において有用な知見が得られた。本稿ではこれについて報告する。

キーワード：JGSS，ISCO 自動コーディングシステム，サポートベクターマシン

1. はじめに

本稿の目的は、国際的な職業分類である ISCO（International Standard Classification of Occupation）のコーディングに対する自動化システムを検討する中で、SSM および JGSS という 2 種類の現実のデータセットを用いて行った実験により得られた結果について、特に分類の精度に注目して報告することである。

社会調査において重要な変数である「職業」の情報は、自由回答で収集される場合が多いが、統計処理を可能にするためにはコード化（職業コーディング）を行う必要がある。職業コーディングは、より詳細には、自由回答である「仕事の内容」や「従業先事業の種類」と、選択回答である「従業上の地位」や「役職」など職業に関する複数の回答を総合的に判断し、数百種類の職業コードから該当するコードを 1 つ選んで付与するという煩雑な作業である（原・海野 1984）。職業コーディングにおいて用いられるコード体系は、各国における職業の概念の違いによりさまざまに異なっており(1)（岡本 2004, 西澤 2006）、また同じ国であっても年度により異なる場合がある(2)。我が国においては、1988 年に国際労働機関（ILO）で改訂された「ISCO（国際標準職業分類）-88 コード」（以下 ISCO コードと略する）（Bureau of Statistics, International Labour Office 2001）の旧版である ISCO-68 コードの流れを組む「日本標準職業分類」（JSCO ; Japanese Standard Classification of Occupation）を基本とし、国勢調査では JSCO に準拠した国勢調査職業分類、社会調査では JSCO を簡略化した「SSM（Social Stratification and Social Mobility）95 職業分類コード」（以下 SSM 職業コードと略する）（1995 年 SSM 調査研究会 1995）が標準的に用いられている。職業コーディングは、このように用いられるコード体系に違いがあっても、テキスト型のデータである自由回答を分類する作業であるために、コーダにかかる負担が大きく、多大な労力と時間を要するという問題をもつ（高橋 2000, 盛山 2004）。

近年の国際比較研究の活発化に伴い、最近では、国内での標準といえる SSM 職業コードだけでなく、国際的な標準である ISCO コードの付与が要請される場合が増えてきた。ここで、2 つのコードについて簡単に説明する。ISCO コードは各桁が分類上の意味を持つ 4 桁の「桁別分類コード」で、大分類 10 個、亜大分類 28 個、中分類 116 個、小分類 390 個からなる。分類の単位を、個人の遂行する課業（task）とそれにともなう責務（duty）からなる職務（job）に置いている。一方、SSM 職業コードは 3 桁の「順番コード」で、小分類 196 個からなる。ISCO コードは ISCO-68 コードと異なり、新たに職務の遂行に必要な「技能度（skill level、skill specialization）」（＝教育・職業資格）（田辺 2008）を新たに採用しており、SSM 職業コードと ISCO コードの間に単純な対応関係を見いだすことが困難な場合が生じている(3)。したがって、両方のコードが要請される場合は、同一のデータに対してダブル・コーディングを実施する必要があり、コーダの負担や作業時間の問題がこれまで以上に増大する。

実は、職業コーディング自体に内在する作業コストの問題は特に海外で強く認識されており、統計局レベルで自動化に向けての取り組みが行われてきた。例えば、オーストラリア、米国、カナダやフランスなどで自動コーディングの方法が提案されている(4)（Kunz 2003, Creecy et al. 1992, Riviere 1994, 岡本 2004）。ただし、その方法はいずれもデータベースにおける検索の域を出ず、単語や文字単位（場合によっては、2-gram(5) や 3-gram を単位とする）の完全一致による方法を採用している。これに対して、我が国では、2 節で述べるように、SSM 職業コーディングを対象として、表層的ではあるが「意味」まで考慮した自然言語処理に基づいたルールベース手法や、これを機械学習と組み合わせた手法による自動化システムが開発され、利用されている（高橋 2001; 2002; 2003; 2004; 2005, Takahashi et al. 2005 他）。いずれの方法がとられるにせよ、コーダに対して、今後もコンピュータによる有効な支援が検討される方向にあることが予想されよう。

以上を踏まえると、ISCO コーディングに対しても、SSM 職業自動コーディングシステムと同様に、コーダを支援するための自動コーディングシステムの構築を検討することが有効であると考えられる。しかし、ISCO コーディングに、SSM 職業自動コーディングシステムと同様のアルゴリズムの適用を考えた場合に問題となるのは、SSM 職業自動コーディングシステムで効果のあったルールベース手法が、ISCO コーディングには存在しないことである。したがって、ISCO 自動コーディングにおいても、

ルールベース手法を新たに開発することが必要になるが、これは次の 2 つの理由により行わないことにした。まず、SSM 職業コードの体系以上に複雑な構成をもつ ISCO コード体系に対するルール辞書を人手により作成することは、膨大な作業になることが予想されることである。もう一つは、これと関連するが、職業コーディングで利用される職業コード体系は細かい点も含めると変更されることが多く、今後もその可能性があることを考慮すると、変更のたびにルール辞書のメンテナンスを行うことは大変な手間を要することである。したがって、ISCO 自動コーディングシステムの開発にあたっては、ルールベース手法との組み合わせ手法を参考にしながらも、機械学習に限定した検討を行う方が効果的であると判断できる(6)。

しかし、機械学習にはルールベース手法とは別種の問題がある。すなわち、大量の訓練データ（学習に用いるための分類クラス付きデータ。ここでは ISCO コードが付与されたデータを指す）を必要とすることで、訓練データ数のサイズが小さい場合には高い分類精度を期待できない。特に ISCO コーディングの場合は、コード（機械学習ではクラスとよぶ）が 400 個近くもあるため、各々のクラスを学習するのに充分なサイズの訓練データを用意するのは容易ではない。そこで、この問題を回避するために、高橋（2007; 2008）では、ISCO 自動コーディングにおいて学習に用いる素性として何が有効であるかという「素性選択」について検討を行うために、2003 年 SSM データセット（767 サンプル）を用いた実験を行った（関連研究 2.2 を参照のこと）。また、高橋（2010）では、対象を ISCO 自動コーディングに限定せず、機械学習全般に対して汎用性のある新しいアンサンブル学習(7)手法についての研究を行っている。

このような状況の中で、ISCO が付与された 2005 年 SSM 日本調査（以下、2005SSM データセットと略する）に加えて、今回、JGSS データセットの利用が可能になった。両者はいずれも我が国における代表的な大規模調査により収集されたデータであり、かつ調査後に付与された職業コードの信頼性が高いため、ISCO 自動コーディングにおける訓練データとして貴重である。両者を併合すれば、正確性のある訓練データのサイズが拡大することになり、実際問題として効果的である。しかし、両者は類似しているものの、調査主体が異なっており、同じデータセットであるとはいえない。実際、高橋（2008）では、SSM 職業自動コーディングにおいては、2005SSM データセット（本人現職）に対して JGSS データセットを訓練データとした場合に、JGSS データセットに対する場合より分類精度が低下したことが報告されている(8)。また、より厳密にいえば、JGSS データセットに限っても、職業データの種類（本人現職か配偶者現職かなど）やコーディングの時期（言い換えれば、正解コードの決定者）が異なる場合は、全く同一の性質をもっているわけではないが、これについては、SSM 職業自動コーディングにおいて用いた職業データの種類やコーディングの時期が異なっていても、訓練データのサイズが大きいほど分類精度が向上することが観察されており、同一のデータセットとみなすことが可能であると判断できる（高橋 2005）。

SSM 職業自動コーディングにおける以上の結果を考慮すると､ISCO 自動コーディングにおいては、少なくとも、調査主体が同一の訓練データを用いることが望ましいことは明らかである。しかし、我が国では ISCO コーディングが実施されてからの期間が短いために、訓練データの蓄積が不足しているという事情があり、この条件が満たされる可能性は高くない。そこで、現実的な問題として、ISCO 自動コーディングにおいて、調査主体の異なるデータセットを併合した場合の状況について明らかにしておくことが必要であろう。また、前述した素性選択の問題についても、これまでは小規模な実験から得られた結果しかないために、そこでの結論が一般化できるかどうかについての確認を行っておくことも必要である。本稿ではこの 2 つの目的のために、2005SSM および JGSS データセットを用いた実験を行い、結果を報告する。

以下、次節で、本稿で行う実験の背景となる ISCO 自動コーディングの分類精度向上に関連する研究について述べた後、3 節で実験を行い、4 節で結果と考察を述べる。最後に 5 節でまとめる。

2. 関連研究

ここでは、まず、職業コーディングの自動化という点で参考にできるSSM職業自動コーディングシステムについて述べる。次に、このシステムで最も有効性の高かったアルゴリズムをISCO自動コーディングに対して拡張する方法を検討した研究について述べる。

2.1 SSM職業自動コーディングシステム

SSM職業自動コーディングとして開発されたシステムには、次の3種類がある。すなわち、(1)ルールベース手法[9]（高橋2000)、(2)機械学習[10]（高橋他2004)、(3)ルールベース手法と機械学習を組み合わせた手法（4つ）[11]（高橋他 2005, Takahashi et al. 2005）である。このうち、分類精度が最も高かったのは、ルールベース手法と機械学習を組み合わせた手法で、中でも、ルールベース手法の結果を機械学習の素性として追加し学習させる方法（以下add-code法と呼ぶ）が最もよい結果であった。add-code 法は、実験後に実際に適用された調査（JGSS-2003）において、80.7%というさらに高い分類精度を示した[12]（高橋2004)。ISCO自動コーディングに対してもこのアルゴリズムを適用できれば効果的であるが、1 節で述べたように、新たにルールベース手法を開発することはコスト面で問題があるために行わない。

2.2 ISCO自動コーディングにおける機械学習

2.2.1 add-code法の拡張

ルールベース手法を機械学習に組み入れたadd-code法は、そのままの形ではISCO自動コーディングに適用できない。しかし、ここで、ルールベース手法を一種の「分類器」であると考えれば、他の機械学習分類器による出力を素性として利用する機械学習法、すなわち、カスケード（元田他2006）の一種であるとみなすことも可能である。さらに、この考え方を拡張すれば、ルールベース手法による結果でなくても、何らかの出力結果として、例えば、別のシステムであるSSM職業コーディングの結果を、ISCO自動コーディング素性として追加する方法もあり得る。このとき、厳密には、add-code法では、2つの分類器で用いられたクラスがともに同じSSM職業コード体系であったのに対して、この場合には、職業というドメインは同じものの職業コード体系が異なるという問題がある。しかし、add-code 法で利用された素性が同じクラス同士の直接的な知識であるとするならば、この方法では、利用される素性のクラスは異なるが、何らかの関連をもつ間接的な知識であると考えることもできるため、不都合はないと判断した。

さて、素性として追加するSSM職業コーディングの結果としては、自動化システムにより予測されたコード（以下予測 SSM 職業コードと略する）および最終的に付与された正解のコード（以下正解SSM職業コードと略する）の2種類が想定できる。高橋（2008）では、2003年SSMデータセットを用いた実験を行ったが、予測SSM職業コードや正解SSM職業コードの追加は、分類精度をそれぞれ3.3%、9.6%向上させた。予測SSM職業コードより正解SSM職業コードの方がよい結果が得られたのは当然であるが､正解SSM職業コードはSSM職業コーディングが完了するまでは決定できないため、予測SSM職業コードしか利用できない状況もあり得るものと思われる。

2.2.2 学歴の追加

1 節で述べたように、ISCO コードでは SSM 職業コードと異なり、分類の基準として技能度が採用されている。高橋（2008）は、職業情報に技能度を測る変数が直接には含まれていない点に注目し、「学歴」で代用することを検討した。実際、田辺他（2008）においても、技能度と学歴に対して、表1に示すような対応付けが考えられている。

表 1　技能度（スキルレベル）と学歴の対応（田辺他（2008）より掲載）

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

しかし、実験の結果、7 段階にまとめた学歴を素性に追加すると、分類精度が 1.3%低下し、5 段階にまとめた場合も追加しない場合との違いが認められなかった（高橋 2008）。この結果、予想に反して、素性として学歴を用いることは有効ではなかったが、これが ISCO 自動コーディングにおける一般的な結論であるのか否かについて、本稿で確認しておく必要がある。

3. 実験

3.1 実験の目的

本稿における実験の目的は、現在の状況で実際に ISCO 自動コーディングを行う場合に問題となる次の 2 点、すなわち、(1) 有効な素性の確認、(2) 訓練データに SSM と JGSS という異なるデータセットを併用した場合の効果についての実際的な知見を得ることである。

実験目的 (1) のために本稿で用いる素性は次の通りである。まず、自動コーディングのための基本的な素性（以下、基本素性と略する）としては、「仕事の内容」「従業先事業の種類」「従業上の地位・役職」を用いた。追加を検討する素性は、学歴および正解 SSM 職業コードの 2 つに絞った（予測 SSM 職業コードは検討しなかった）。したがって、実験で用いる素性としては次の 4 通り（①～④）を想定した。なお、基本素性として職業情報のすべてを用いたわけではなく、高橋他（2005）において有効性が認められなかった「従業先事業の規模」は、SSM 職業自動コーディングと同様の場合と同様に今回も用いなかった。

①基本素性（baseline）
- 「仕事の内容」に出現する品詞付き単語（原形）(13)
- 「従業先事業の種類」に出現する品詞付き単語（原形）(13)
- 「従業上の地位・役職」の選択肢（10 種類）

②基本素性＋「学歴」の選択肢（13 種類）
③基本素性＋「正解 SSM 職業コード」（約 200 種類）
④基本素性＋「学歴」の選択肢＋「正解 SSM 職業コード」

3.2 実験設定

3.2.1 分類器

分類器はサポートベクターマシン（SVM）を用いた(14)。ただし、SVM は本来 2 値分類器であるために、one-versus-rest 法（Kressel 1999）により多値分類器に拡張した。また、SVM におけるカーネル関数は線形カーネルを用いた。以上の設定は、SSM 職業自動コーディングの場合と全く同様である。

3.2.2 データセットと正解

データセットは、2005SSM および JGSS（JGSS-2006、JGSS-2008、JGSS-2010）の 2 種類を用いた。具体的には、2005SSM では、本人現職、本人初職、配偶者現職、配偶者結婚時職のうちの有職者各 4,133 サンプル、5,542 サンプル、2,915 サンプル、3,499 サンプルの計 16,089 サンプルを用い、JGSS-2006、JGSS-2008、JGSS-2010 では、ISCO コードおよび SSM 職業コードが付与された各 2,224 サンプル（本人現職、配偶者現職）、1,375 サンプル（本人現職）、2,570 サンプル（本人現職、配偶者現職）を用いた。

実験では実際に実施される状況を想定し、これらのデータセットを目的に応じて、単独または組み合わせて用いた（表 2 参照）。表 2 における交差検定とは、機械学習の実験結果の信頼性を高めるための方法で、学習に用いる訓練データと手法を評価するための評価データの分割を行う際に、実験に用いるデータ全体から両者をランダムに選んで複数個のセットを生成し、セットごとに訓練事例と評価事例を用いた実験を行い、得られた結果の平均値を最終的な結果とするものである。例えば、5 分割交差検定では、データ全体の 4/5 を訓練事例、1/5 を評価事例とするセットを 5 通り生成し、計 5 回の実験による結果の平均値が最終結果となる。実験 2 および実験 3-1、3-2 の実験目的は（2）であるが、いずれも、3.1 で述べた 4 種類の素性による実験を行って、実験目的（1）の確認を行った。

なお、本稿における正解とは、調査後に実施された職業コーディングにより付与されたコードである。

表 2　実験目的と実験名および用いたデータセット

<table>
<tr><th>実験目的</th><th>実験名</th><th>訓練データ</th><th>評価データ</th></tr>
<tr><td rowspan="2">（1）</td><td>実験 1-1</td><td colspan="2">2005SSM、JGSS-2006、JGSS-2008、JGSS-2010
＊訓練データと評価データの分割はいずれも 5 分割交差検定によった</td></tr>
<tr><td>実験 1-2</td><td colspan="2">2005SSM & JGSS-2006、2005SSM & JGSS-2006 & JGSS-2008、
2005SSM & JGSS-2006 & JGSS-2008 & JGSS-2010
＊訓練データと評価データの分割はいずれも 5 分割交差検定によった</td></tr>
<tr><td rowspan="4">（2）</td><td>実験 2</td><td>2005SSM（固定）</td><td>2005SSM
JGSS-2006
JGSS-2008
JGSS-2010</td></tr>
<tr><td>実験 3-1</td><td>JGSS-2006、JGSS-2006 & JGSS-2008</td><td rowspan="2">JGSS-2010
（固定）</td></tr>
<tr><td>実験 3-2</td><td>2005SSM & JGSS-2006、2005SSM & JGSS-2006 & JGSS-2008</td></tr>
</table>

3.2.3 評価尺度

評価尺度は、正解率（＝正しく分類できた事例数／全事例数）を用いた。本稿では、クラス別に算出した個々のミクロな正解率ではなく、全クラスから算出したマクロな正解率を用い、この値が高いほど有効であるとした。今回はすべての事例にコードが付与されているため、正解率は、文書分類で評価尺度として用いられる「分類精度（classification accuracy）」と「再現率（recall）」のいずれとも一致する。したがって、以下では、これらの概念をすべて正解率なる用語で統一して述べる。

4. 結果と考察

表 2 に示した実験ごとに結果を述べて考察を行う。なお、以下の表で示す①～④は、3.1 で述べた素性の組み合わせ方の番号を示す。

4.1 素性選択の確認

ここでは、実験 1-1 および実験 1-2 により、有効な素性の組み合わせを調査した結果を報告する。**［実験 1-1］**まず、単独のデータセットごとに、有効な素性の組み合わせを調査した結果、いずれも基本素性に正解 SSM 職業コードを追加した場合が最もよい結果を示した（表 3 参照）。また、ここでも 2.2.2 で述べた実験と同様に、学歴は有効ではなく、基本素性のみの場合よりも正解率が低下した場合もあった。なお、素性のどの組み合わせにおいても、2005SSM データセットにおける正解率が最も高かったが、この理由は、訓練データのサイズの効果によるものであると考えられる。

表 3　データセット別有効な素性（太字は 4 種類の素性中最も高い正解率を示す）

	訓練&評価データ			
素性の組み合わせ方	2005SSM (15,271)	JGSS-2006 (1,779)	JGSS-2008 (1,086)	JGSS-2010 (2,056)
①（baseline）	0.6834	0.5323	0.5356	0.6051
②	0.6832	0.5323	0.4945	0.6051
③	**0.7448**	**0.5863**	**0.6571**	**0.6882**
④	0.7425	0.5863	0.6531	0.6882

データセット名の下段の数値は、訓練データのサイズを示す。

[実験 1-2] 次に、実際に適用する場面を想定し、2005SSM データセットに JGSS データセットを順次追加して生成したデータセットにより、有効な素性の組み合わせを調査した（表 4 参照）。ここでも、すべての場合で、基本素性に正解 SSM 職業コードを追加した場合が最もよい結果を示した。2005SSM データセット単独の場合は、基本素性のみの場合より 6.1%の上昇であったが、追加する訓練データが増えるほど、すなわち訓練データのサイズが大きくなるにつれて正解率が向上し、すべての JGSS データセットを追加した時点では 17.4%まで上昇した。

学歴に関しては、実験 1-1 と同様に、13 種類のデータを 2.2.2 で述べた実験のようにまとめずにそのまま用いたにもかかわらず、基本素性のみの場合よりわずかに高い値となった。この理由は、訓練データのサイズが実験 1-1 の場合より大きくなったことで、ノイズの影響が減少したためであると思われる。2.2.2 で述べた実験では、7 段階にまとめても正解率が低下し、5 段階でも高くならなかったが、この理由は、素性の変動範囲に比べて訓練データのサイズが非常に小さかったためではないかと思われる。以上より、学歴は、素性の取り得る範囲を減らす状況下で（例えば表 1 に示した 4 段階にまとめる）、訓練データのサイズが大きい場合には、有効な素性となる可能性がある。これについては今後の課題としたい。

表 4　併合したデータセット別有効な素性（太字は 4 種類の素性中最も高い正解率を示す）

	訓練&評価データ		
素性の組み合わせ方	2005SSM & JGSS-2006 (17,050)	2005SSM & JGSS-2006 & JGSS-2008 (18,136)	2005SSM & JGSS-2006 & JGSS-2008 & JGSS-2010 (20,192)
①（baseline）	0.6780	0.6308	0.5513
②	0.6785	0.6342	0.5536
③	**0.7368**	**0.7156**	**0.7252**
④	0.6833	0.6849	0.6851

データセット名の下段の数値は、訓練データのサイズを示す。

4.2 訓練データに異なるデータセットを併用した場合の効果

ここでは、実験 2、実験 3-1 および実験 3-2 により、JGSS データセットを評価データとした場合に、調査主体の異なる 2005SSM データセットを訓練データとして用いることによる効果について調査した結果を報告する。

[実験 2] まず、2005SSM データセットの有効性を確認するために、2005SSM データセットを単独で訓練データとしたときの JGSS データセットにおける正解率を調査した（表 5 参照。比較のため、2005SSM データセット自身に対する正解率も示した）。表 5 に示すように、いずれの場合も 2005SSM データセットの利用は有効性が低く、2005SSM データセット自身に対する場合より平均で約 6.6%～

10.3%低かった。

JGSS の 3 つのデータセットの中で最も正解率が高かったのは JGSS-2008 で、2005SSM データセットが最も効果的であったといえる。また、素性の組み合わせ方による正解率の高低の傾向が最も類似しているのは JGSS-2006 であった。一方で、JGSS-2010 は、データセットの性質が最もシンプルに現れる基本素性のみによる場合の正解率がきわめて低かった。4 つのデータセットに出現した ISCO コードを調査すると、JGSS-2010 以外の 3 つはいずれも、我が国の事情に合わせた新規の ISCO コード（「3418」「5249」「5164」）（田辺他 2008)）が付与されていたが(15)、JGSS-2010 では、これらのコードはなく、これらに対応する既存のコード（「3410」「3412」「0110」）のみが付与されていた。これより、JGSS-2010 は他の JGSS データセットと同一には扱えず、また 2005SSM データセットとも異なる性質をもつ。したがって、これ以降の実験においては JGSS-2010 を評価データとして扱うが、これは、2005SSM や JGSS 以外の一般的な調査を想定した実験であるとみなすことも可能である。ここで興味深いのは、JGSS-2010 は、基本素性のみの場合の正解率が最も低いが、正解 SSM 職業コードを追加すると最も高くなることである。これより、正解 SSM 職業コードは、データセット間の性質の違いを吸収する効果があるのではないかと考えられる。

表 5　2005SSM データセットを訓練データとしたときの各 JGSS データセット別正解率
（太字は素性の組み合わせ別に各データセットの中で最も高い正解率を示す）

訓練データ：2005SSM（16,089）	評価データ			
素性の組み合わせ方	2005SSM（参考値）	JGSS-2006	JGSS-2008	JGSS-2010
①	0.6834	**0.5899**	0.5770	0.5697
②	0.6832	0.5863	**0.5925**	0.5805
③	0.7448	0.6093	0.6699	**0.6997**
④	0.7425	0.6057	**0.7015**	0.6482
①～④の平均	0.7010	0.5978	**0.6352**	0.6245

データセット名の下段の数値は、訓練データのサイズを示す。

［実験 3-1］次に、2005SSM データセットを利用せずに、JGSS データセットを単独で訓練データとしたときの正解率を調査した（表 6 参照)。このとき、実際の適用を想定し、過去のデータセット（JGSS-2006 や JGSS-2008）を訓練データと考え、最も新しい JGSS-2010 を評価データとした。表 6 に示すように、同じ JGSS 内であれば、単純に訓練データのサイズが大きい方（JGSS-2006 と JGSS-2008 の両方を用いる場合）がよい結果であり、特に、正解 SSM 職業コードを追加した場合は最もよかった。しかし、正解率が 61.5%という値は、システムの実用化の点からは明らかに低く、実際、訓練データとして JGSS データセットを全く利用しない場合（表 5 最右列参照）よりも約 8.5%低いという有様である。これは、訓練データのサイズの問題とデータセットの異なり方のどちらが優勢なのかという問題に起因するのではないかと考えられるため、当面は、調査主体が違っていても、はるかにサイズの大きなデータセット（例えば 2005SSM データセット）が利用できる場合は、これを訓練データとする方が効果的であると考えられる。

表 6　JGSS データセットを訓練データとしたときの正解率（太字はすべての中で最も高い正解率を示す）

評価データ：JGSS-2010	訓練データ	
素性の組み合わせ方	JGSS-2006 (2,224)	JGSS-2006 & JGSS-2008 (3,581)
①	0.5148	0.5529
②	0.5148	0.5502
③	0.5852	**0.6148**
④	0.5852	0.6109

データセット名の下段の数値は、訓練データのサイズを示す。

[実験 3-2] 最後に、本稿の目的である 2005SSM データセットと JGSS データセットを併合して訓練データを生成したときの正解率を調査した（表 7 参照）。ここでも実際の適用を想定し、最も新しい JGSS-2010 を評価データとし、過去のデータセットを訓練データとして扱った。これまでの結果から、訓練データとして 2005SSM データセット、JGSS-2006、JGSS-2008 を併合し、素性として正解 SSM 職業コードを追加した場合が最もよい結果になることが予想されたが、表 5、表 7 から明らかなように、この値は、訓練データとして 2005SSM データセットを単独で利用した場合の値を約 4.6%も下回る結果となった。ただし、これ以外の素性の組み合わせ方では、これらのデータセットを併合した方がよかった。なぜ、正解 SSM 職業コードを素性として追加した場合のみ訓練データのサイズが有利に働かなかったのかということについては、今後の課題とするが、いずれにしても、現在の状況においては、訓練データをどのように生成すればよいかという問題については、表 5～表 7 に示した結果から、2005SSM データセットの援用（利用できる素性により単独または JGSS データセットと併用）が有効であるといえる。

表 7　2005SSM および JGSS データセットの両方を訓練データとしたときの正解率
（太字はすべての中で最も高い正解率を示す）

評価データ：JGSS-2010	訓練データ	
素性の組み合わせ方	2005SSM & JGSS-2006 (18,313)	2005SSM & JGSS-2006 & JGSS-2008 (19,670)
①	0.5724	0.6016
②	0.5856	0.5969
③	0.6412	**0.6533**
④	0.6191	0.6521

データセット名の下段の数値は、訓練データのサイズを示す。

以上の実験を通じて、現状において、ISCO 自動コーディングにおける訓練データとして利用可能なデータセット（約 20,000 サンプル）に関して、次のような実際的な知見が得られた。(1) 素性選択：基本素性と正解 SSM コードの併用である。学歴は正解率を下げるとは結論できず、より大きなデータセットで再度、確認を行う必要がある。(2) 訓練データ：素性として正解 SSM コードが利用できる場合は、2005SSM データセットのみを用いる。正解 SSM コードが利用できない場合は、2005SSM データセットと JGSS データセットの両方を用いる。いずれの場合も、JGSS データセットのみを訓練データとするのは有効ではない。

ところで、1 節で述べたように、自動コーディングシステムの目的はコーダの支援であり、システムの結果がそのまま最終的な決定となるわけではない。しかし、このことを考慮しても、現時点で得られる正解率の低さでは、有効な支援が期待できないものと思われる。そこで、コーダに対するヒン

トとして、複数個の結果を提示することを考えてみる。図 1～図 3（図の X 軸は利用する順位、Y 軸は正解率を示す）は、評価データを JGSS-2010 とし、訓練データをそれぞれ JGSS データセット単独、2005SSM データセット単独、2005SSM と JGSS のデータセットの併合とした場合の、第 1 位から第 5 位までの結果[16]による正解率を示したものである。これにより、複数個の結果を利用すれば、正解率が高まることがわかる。ただし、不要に多くの結果を提示することはコーダの混乱を招くため、正解の可能性が低い結果は提示しないなどの工夫も必要であり、この問題については、現在研究中である（Takahashi et al. 2008）。

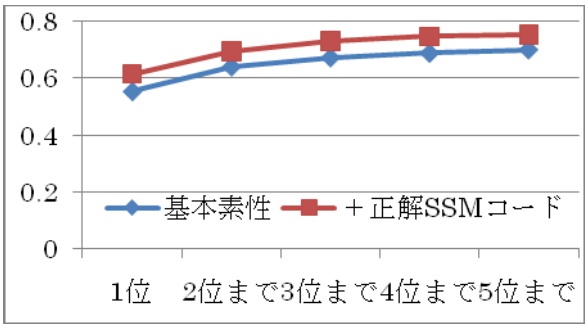

図 1　訓練データ：JGSS-2006 & JGSS-2008

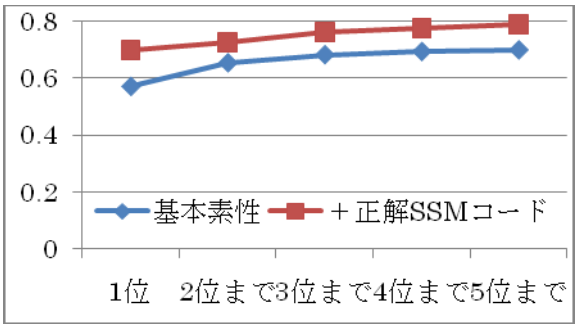

図 2　訓練データ：2005SSM 単独

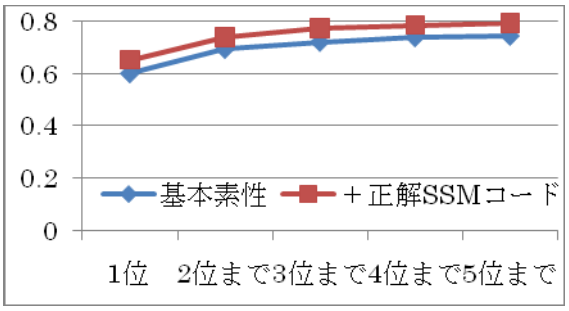

図 3　訓練データ：2005SSM & JGSS-2006 & JGSS-2008

5. おわりに

本稿では、ISCO 自動コーディングにおいて、現時点で利用可能な SSM データセットと JGSS データセットの両方を用いた実験を行い、そこで得られた実際的な知見について報告した。この中で有効性の高かった最終的に決定された SSM 職業コードを素性として追加できる場面としては、SSM 職業コードが付与されている過去のデータセットに対して、新しく ISCO コードの付与を行う場合であろう。一方、新規に職業コーディングを行う場合は、最終的な SSM 職業コードが決定される前に ISCO コーディングが行われる可能性が高いため、有効性の点でやや劣るが、ルールベース手法によりあらかじめ予測された SSM 職業コードの利用が望まれる。いずれの場合も、現段階では ISCO コードが付与されたデータセットの量が十分ではないために、自動コーディングシステムが出力する第 1 位の結果しか利用されない場合は正解率が高くない状況であり、必要に応じて第 2 位以下の結果も利用することが勧められる。ただし、この問題は、今後、ISCO コーディングの実施が増えれば、ISCO コードが付与されたデータセットの蓄積が進むため、徐々に改善されることが期待できよう。

今後の課題は、利用可能な訓練データのサイズの制約から今回の実験では確認できなかった点、すなわち素性選択における学歴の有効性について、サイズが増大した時点で同様の実験を行うことである。また、本稿で得られた結論を実際の ISCO 自動コーディングシステムに活用することも重要である。その一環として、現在、Web 公開を構想中の「SSM 職業コードおよび ISCO コードに対する自動コーディングシステム」に本稿での結論を反映させる予定である。

[Acknowledgement]
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。
2005 年 SSM 調査データの利用に関して、2005 年 SSM 調査研究会の許可を得た。

[注]

（1）例えば、英国では SOC 2000、オーストラリアでは ASCO2、米国では SOC、カナダでは NOC-S2001 なる職業コード体系が用いられており、小分類はそれぞれ順に 353 個、340 個、449 個、520 個である。このうち、ASCO2 や SOC ではさらに細分類があり、それぞれ 986 個、821 個である。
（2）例えば、米国センサスでは、1970 年と 1980 年で異なる職業コード体系が用いられた（Rubin 2004）。
（3）例えば、『職業分類における SSM95 と ISCO-88 の対応』（2005 年 SSM 調査研究会職業分類タスクグループ 2004）によれば、「2003 年仕事と暮らしに関する全国調査」において 1 対 1 に対応する事例は約 30～40%程度であったとの報告がなされている。このような事情により、「2005 年 SSM 社会階層と社会移動日本調査」（以下「2005 年 SSM 日本調査」と略する）では、ISCO コードを SSM 職業コードから変換するのではなく、新たに ISCO コーディングを行って付与することにした。SSM 職業コードと ISCO コードの対応関係については鹿又他（2008）に詳しい。
（4）例えば、オーストリアではシソーラスベースの Precision Data、米国では AIOCS や PACE、カナダでは ACTR、フランスでは SICORE と呼ばれる自動コーディングシステムが提案されており、英国、イタリア、アイルランドにおいても検討が行われている（Keogh 1998）。
（5）例えば、「work at the office」という回答の場合、単語単位の 2-gram では、「work at」「at the」「the office」が処理の単位となり、文字単位の 2-gram では、「wo」、「or」、・・・、「ic」、「ce」が処理の単位となる。
（6）今回の ISCO コーディングに限らず、今後のシステム開発もこの方針に基づいて行う予定である。
（7）アンサンブル学習法とは、機械学習において複数の分類器を構築することで分類精度を高める手法の総称である（元田他 2006）。
（8）訓練事例として、評価事例とやや性質が異なる JGSS データ（約 34,521 サンプル）を用いたため、訓練事例数が JGSS-2003 より約 15,000 サンプル多かったにもかかわらず正解率は 8.3%低かった。
（9）職業を表現する手がかりとして動作を表す述語に注目し、文中の名詞を述語との関係で捉える「格フレーム」の概念を利用して、職業の定義内容を表現するルールを生成した。この他に、コード・ブック（1995 年 SSM 調査研究会 1996，西村・石田 2001，石田・三輪 2006）に記述された内容やコーダが判断する際に用いる知識も可能な限りルール化されている。また、これらのルールと回答の意味的な一致を捉えるために、両者に出現する述語と名詞をそれぞれシソーラスにより拡張した。ルールベースによる方法による自動コーディングの結果は、コーダが利用しやすいように表計算用ソフトにより表示する機能が追加され、ROCCO（Rule-based OCcupation COding）システムと名付けられた。ROCCO システムは、JGSS を始めとする調査で利用され、一定の評価を得た。
（10）ルールベースによる方法は人間が結果を理解しやすい反面、準備されたルールにマッチした事例しかコードを決定できないことや、ルール辞書のメンテナンスに手間がかかるという問題がある。これに対して、機械学習による方法は、自由回答を非常に短い文書と捉えることで、文書分類と同様の分類方法を適用しようとするもので、自由回答に出現する品質付きの単語を分類器が学習する素性として利用することを想定する。機械学習にはいくつかの方法があるが、多くの文書分類タスクにおいて精度の高さが評価されているのはサポートベクターマシン（SVM）（Vapnik 1998, Sebastiani 2002）であるため、SVM を用いた。
（11）機械学習による方法はルールベース手法を上回る性能を示したが、コーダと同等の性能を目指すために、ルールベース手法のもつ貴重な「知識」を活かす方法が検討された。
（12）この理由は、訓練事例が約 13,300 サンプルから約 20,000 サンプルに増えたためであると考えられる。
（13）形態素解析用ソフトとして、京都大学長尾研究室で開発された JUMAN（黒橋・長尾 1998）を用いた。

（14）http://chasen.org/~taku/software/TinySVM/ を利用した。このため、3.3.1①で述べた品質付き単語は離散値に変換したものを素性とした。

（15）この 3 つのコードが各データセットで占める割合は、いずれも合計で 1%程度であった。

（16）SVM では、事例に対して予測した分類クラスに付随して事例の分離平面からの距離（スコア）を出力するが、多値分類の場合は、スコアの大きさにしたがって予測した分類クラスが順位付けられる。

[参考文献]

1995 年 SSM 調査研究会, 1995,『SSM 産業分類・職業分類（95 年版）』1995 年 SSM 調査研究会.

1995 年 SSM 調査研究会, 1996,『1995 年 SSM 調査コード・ブック』1995 年 SSM 調査研究会.

2005 年 SSM 調査研究会職業分類タスクグループ, 2004,『職業分類における SSM95 と ISCO-88 の対応』.

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

Creecy, R. H., Mas, B. M., Smith, S. J., and Waltz, D. L., 1992, "Trading Mips and Memory for Knowledge Engineering," *Communication of the ACM* 35(8): 48-63.

Dumais, S., Platt, J., Hecherman, D., and Sahami, M., 1998, "Inductive Learning Algorithms and Representations for Text Categorization," *Proceedings of the ACM-CIKM*98: 145-155.

Gillman, D. W., and Appel, M. V., 1999, "Developing an Automated Industry and Occupation Coding System for CENSUS 2000," *2000 Proceeding of the American Statistical Association Annual Meeting*, Government Statistics Section.

原純輔・海野道郎, 1984,『社会調査演習』東京大学出版会.

石田浩・三輪哲, 2006,『2005 年 SSM 調査 SSM95 職業・産業コーディングマニュアル改訂版』東京大学社会科学研究所付属日本社会研究情報センター.

Joachims, T., 1998, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *Proceedings of the European Conference on Machine Learning*: 137-142.

鹿又信夫・田辺俊介・竹ノ下弘久, 2008,「SSM 職業分類と国際的階層批評：EGP 階級分類・SIOPS・ISEI への変換」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 平成 16～19 年度科研費特別推進研究「現代日本階層システムの構造と変動に関する総合的研究」報告書』2005 年 SSM 調査研究会: 69-94.

Keogh, G., 1998, "Automatically Coding Occupation Description from the 1996 Census of Population of Ireland," *Technical report in Central Statistic Office (CSO)*.

Kunz, C., 2003, *CENSUS: OCCUPATION (Census Paper No.03/06)*, Australian Bureau of Statistics.

黒橋禎夫・長尾眞, 1998,『日本語形態素解析システム JUMAN version 3,61』京都大学情報研究科.

Kressel, U., 1999, "Pairwise classification and Support Vector Machines," Scholkopf, B., Burgesa, C. J. C., and Smola, A. J. [eds.], *Advances in Kernel Methods Support Vector Learning*, The MIT Press, 255-268.

元田浩・津本周作・山口高平・沼尾正行, 2006,『データマイニングの基礎』オーム社.

西村幸満・石田浩, 2001,『SSJ Data Archive Research Paper Series JGSS-2000 調査 職業・産業コーディングインストラクション』東京大学社会科学研究所付属日本社会研究情報センター.

西澤弘, 2006,『労働政策研究報告書 No.57 職業紹介における職業分類のあり方を考える「労働省編職業分類」の改訂に向けた論点整理―』独立行政法人労働政策研究・研究機構.

岡本政人, 2004,「国内外における統計自動格付法の研究動向」『製表技術参考資料 2』独立行政法人統計センター: 46-77.

Riviere, P., 1997, "SICORE - general automatic coding system," *Statistical Data Editing Vol.2 Methods and Techniques*, United Nations Statistical Commission and Economic Commission for Europe, 222-231.

Rubin, D. B., 2004, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Hoboken New Jersey.

Sebastiani, F., 2002, “Machine Learning Automated Text Categorization,” *ACM Computing Surveys* 34(1): 1-47.

盛山和夫, 2004,『社会調査演習』東京大学出版会.

高橋和子, 2000,「自由回答のコーディング支援について―格フレームによる SSM 職業コーディングシステム―」『理論と方法』15(1): 149-164.

高橋和子, 2001,「自由回答のコーディング自動化システム―「健康と階層」調査における職業コーディング―」(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)研究成果)『敬愛大学国際研究』8(1): 31-52.

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』1: 171-183.

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: 179-191.

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: 163-174.

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―職業コーディングの自動化の方法―」『理論と方法』15(1): 177-196.

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』12(2): 3-24.

Takahashi, K., Takamura, H., and Okumura, M., 2005, “Automatic Occupation Coding with Combination of Machine Learning and Hand-Crafted Rules,” Bao, H. T., David, C., and Huan, L. [eds.], *Advances in Knowledge Discovery and Data Mining Proceedings Series: Lecture Notes in Computer Science Subseries: Lecture Notes in Artificial Intelligence* 3518: 269-279, Springer-Verlag Berlin Heidelberg.

高橋和子, 2008,「機械学習による ISCO 自動コーディング」『2005 年 SSM 調査シリーズ 12 社会調査における測定と分析をめぐる諸問題(前田忠彦編)平成 16～19 年度科研費特別推進研究「現代日本階層システムの構造と変動に関する総合的研究」報告書』2005 年 SSM 調査研究会: 47-68.

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』15(2): 3-38.

Takahashi, K., Takamura, H., and Okumura, M., 2008, “Direct estimation of class membership probabilities for multiclass classification using multiple scores,” *Knowledge and Information Systems (KAIS)*, 19(2): 185-210, Springer-Verlag, London.

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』https://kaigi.org/jsai/webprogram/2010/pdf/260.pdf

田辺俊介, 2008,「SSM 職業分類と ISCO-88 の比較分析」『2005 年 SSM 日本調査の基礎分析-構造・趨勢・方法(前田忠彦編)平成 16～19 年度科研費特別推進研究「現代日本階層システムの構造と変動に関する総合的研究」報告書』2005 年 SSM 調査研究会, 31-47.

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 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.

Vapnik, V., 1998, *Statistical Learning Theory*, John Wiley, New York.