機械学習とルールベースの組み合わせによる職業コーディング

高橋 和子[†]　　高村 大也[††]　　奥村 学[††]
[†]敬愛大学 国際学部　　[††]東京工業大学 精密工学研究所
takak@u-keiai.ac.jp　　{takamura,oku}@pi.titech.ac.jp

1　はじめに

社会調査において, 職業は重要な属性の一つで細かい分類（約 200）を必要とするため (1995 年 SSM 調査研究会, 1995), 調査票に職業コードを提示せずに, 被調査者から自由回答で記述してもらったデータを分析者側で分類することが多い (1995 年 SSM 調査研究会, 1996)(西村 · 石田, 2000). この作業は職業コーディングとよばれ, データの分析前に人手（コーダ）により行われてきたが, サンプル数が多い場合にはコーダの負担が大きく, また, 熟練者でないとコーディングの結果に一貫性が欠如しやすいという問題が存在した.

この問題を軽減するために, 職業の定義文や専門家の知識をルールとしてまとめた辞書を作成しておき, 回答とのマッチングをとって自動的に職業コードを決定するルールベースによるシステムを開発した (高橋, 2000). システムはこれまでに JGSS（日本版 General Social Surveys）[1] を始めとする 6 つの調査で利用され, 正解率は約 60〜70% 程度であったが, コーダの負担を軽減し, 結果における一貫性の保証という点で, 当初の目的を達成できたといえる (高橋, 2002; 高橋, 2003; 高橋, 2001).

しかし, システムには次のような問題が残っていた. まず, 職業に関するデータ（以下, 職業データと呼ぶ）は, 単一の回答ではなく, 自由回答である「仕事の内容」を中心に, 「従業先事業の種類」（自由回答）,「従業上の地位」,「役職」,「従業先事業の規模」（以上は選択肢）を含む一連の回答群から構成されており, これらは総合的に判断されて分類先が決められる (1995 年 SSM 調査研究会, 1996). 従って, そこで用いられる知識をすべてルールとして表現することは非常に困難である. 次に, システムは, 自由回答とルールをいずれも格フレームの形式で表現するが, この形式で表現できない回答もあり[2], データの処理ができない場合がある. さらに, 職業データの中で最も重要な「仕事の内容」に出現する用語や表現の仕方は時代と共に変化しており, システムのもつ辞書やシソーラスの継続的なメンテナンスの手間が必要なことである.

そこで, これらの問題を回避する方法として, テキスト分類の分野に適用される機械学習手法の中で高い分類性能を示すことが知られている (Joachims, 1998)(工藤 · 松本, 2002)(Sebastiani, 2002) サポートベクターマシン（Support Vector Machine, SVM）に注目した. SVM を職業コーディングに適用した結果, ルールベース手法より正解率が高いことがわかった. また, SVM をすでに存在するルールベース手法と組み合わせると, 正解率が向上することも確認された (高橋他, 2004).

本稿では, これらの結果を発展させ, 両者のさらに有効な組み合わせ方を調査する. 学習の際, 訓練データとして, 過去の事例だけでなくこれからコーディングを行う新たな事例の一部をフィードバックすることで正解率の向上が期待される. そこで, この効果についてコーディング量の観点から調査する.

以下, 次節で関連研究について述べた後, 3 節でルールベース手法による方法, 4 節で SVM による方法およびルールベース手法との組み合わせ方について述べる. 5 節で実験結果について考察を行い, 最後にまとめと今後の課題を述べる.

2　関連研究

関連研究のうち, ここでは人手によるルールに基づく方法と機械学習による方法の組み合わせについて述べる.

文献 (Seng-Bae and Byoung-Tak, 2003) では, 韓国語のチャンキングにおいて, まず人手でルールを作成し, このルールに対するエラーリストも事前に作成しておく. 分類の際は, 事例のチャンクタイプをまずルールで決めた後, このエラーリストの中の最も近い事例との類似度を調べ, その値が閾値以上のものに対しては, memory-based learning により決めるという方法を提案している.

[1] http://jgss.daishodai.ac.jp/において, 2 回の予備調査と 2000 年調査と 2001 年調査についてコーディング済みデータの公開を行っている.

[2] 商品名や生産物だけであったり, 職業と直接関係のない内容を含む場合などがある. 例えば, 1995 年 SSM 調査（Social Stratification and social Mobility survey）データの中から無作為に選んだ 1000 サンプルのうち, 約 20% がこれに該当した (高橋, 2000).

提案手法を用いた場合は, それぞれを単独で用いたときより, 正解率が上がったことが報告されている. また, 文献 (Isozaki et al, 2003) では, より信頼できる日本語ゼロ代名詞の解決のために, ヒューリスティックなルールをランク付けして機械学習（SVM）と組み合わせる方法を提案している. 実験の結果, ME 法や SVM はルールによる方法に劣るが, SVM とルールを組み合わせるとこれを上回ることが報告されている.

3 ルールベース手法

ルールベース手法であるシステムの処理を示す.

1. 「仕事の内容」の回答から述語相当語（以下, 述語と呼ぶ）と, その述語の対象格や場所格となる名詞を抽出する.
2. 述語をシソーラスにより拡張し, 職業コードに関する定義文や知識を
   ＜述語, 格, 名詞＞ ⇒ ＜職業コード＞
   の形式のルールとしてまとめた「職業辞書」を検索する.
3. 「職業辞書」に回答とマッチするルールがあれば, その職業コードを付ける. その際, 回答の名詞も, 述語と同様にシソーラスにより拡張される. また, 回答に「職業辞書」が要求する名詞が不足する場合については, 「従業先事業の種類」の回答も調べ, 述語コードが同じ場合にはこの名詞により補う.
4. 職業データの他の情報（「従業上の地位」, 「役職」, 「従業先事業の規模」）のチェックを行って, 最終的な職業コードを決定する.
   ここで適用されるのは, 次のようなルールである.
   ＜職業コード, 従業上の地位, 役職, 従業先事業の規模＞ ⇒ ＜職業コード＞
   この結果, 先に決定された職業コードが変わる場合もある.

システムには 2 種類のルールがあるが, ルールベース手法においては, これらのルールを充実させることが重要である.

4 SVM による方法

SVM は統計的機械学習に基づく 2 クラスのパターン認識手法である. 学習サンプルと分類境界の間隔（マージン）を最大化するという戦略により, ニューラルネットワークなどの従来手法より, 汎化能力が高く, 精度よく評価事例を分離できるという特徴をもつ.

ここでは, まず, SVM を適用する方法について述べた後, ルールベース手法による方法との組み合わせ方について述べる.

4.1 SVM による方法

実験に用いた基本的な素性（以下, 基本素性と呼ぶ）は次の通りである.

- 「仕事の内容」に出現する単語（原形）
- 「従業先事業の種類」に出現する単語（原形）
- 「従業上の地位＋役職」の選択肢（14 種）[3]

SVM は基本的に二値分類器であるため, 今回の職業小分類のような多値分類に適用するには拡張を行う必要がある. 本稿では, one-versus-rest 法を用いて多値分類器へと拡張した. 以下, その他の方法においても同様である.

4.2 ルールベース手法との組み合わせ

ここでは, 次の 4 つについて考える.

1. ルールベース手法が出力した職業コードを素性に追加する（add-code）
2. ルールベース手法でマッチしたルール（2 種類）を素性に追加する（add-rule）
3. ルールベース手法で出力した職業コードおよびマッチしたルールを素性に追加する（add-code-rule）
4. 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 による方法の結果を利用する（seq）[4]

ここで, 4 は, 機械学習による方法同士の組み合わせではないが, 複数の手法をシーケンシャルに適用する, 一種の ensemble learning であると考えることもできる. 同様に, 1 は, ある分類器からの出力結果をもう一方の分類器の素性としている点で, 一種のスタッキングであると考えられる (Wolpert, 1992).

5 実験

5.1 実験方法

実験の主な目的は, 4 節で述べた各手法に対して, “SVM とルールベース手法の有効な組み合わせ方” と “新しいデータのフィードバック効果” を調査することである. 実験に用いたデータセットは, JGSS により 2000 年から 2002 年まで 3 回実施された調査により収集された有職者のデータ（JGSS-2000, JGSS-2001, JGSS-2002 の 3 データ

[3] JGSS では, 「従業上の地位」と「役職」をまとめた選択肢を用いる (大阪商業大学比較地域研究所 · 東京大学社会科学研究所, 2002).

[4] ルールベース手法が職業コードを決定できない場合とは, 未決定のコードを出力する場合と, 複数個の職業コードを出力する場合をいう. その場合に使用される SVM の訓練データとしては, 全事例を用いる方法と, ルールベース手法がコードを決定できない事例だけを用いる方法の 2 種類がある.

セット, 延べ 20,066 サンプル）[5]である．年度ごとの内訳は, それぞれ, 6848 サンプル, 6448 サンプル, 6770 サンプルである．本稿では, JGSS-2000 データと JGSS-2001 データを訓練データ, JGSS-2002 データを評価データとする[6]．また, フィードバック効果の実験では, 上記の訓練データに JGSS-2002 データの一部を加えたものを訓練データ, JGSS-2002 データの残りを評価データとする．なお, SVM のカーネル関数は線型カーネルを用いた．

5.2 結果と考察

まず, 各手法の正解率を示す（図 1）．ここで, C（ソフトマージンパラメータ）は例外的な事例への許容度合いを表し, 値が小さいほど例外的な事例に与えらえる重みが小さくなる．

C=1.0 のとき, 正解率は

add-code-rule>add-code>seq≈add-rule>SVM

の順である．なお, ルールベース手法の正解率は約 66.1%（JGSS-2002 データ）で, SVM よりさらに約 6% 低かった (高橋他, 2004)．C の値が変化しても, 各手法の順位は, seq と add-rule が入れ替わることを除いてほとんど一定である．特に, SVM は C の値に関係なく, ルールベース手法と組み合わせた手法に劣る[7]．

ここで, C の変化に対する正解率の傾向は手法により違いがある．seq は影響を受けないが, SVM は C=1.0 のとき最大で C の値が小さくなるほど低下するのに対し, 両者を組み合わせた手法はいずれも全く逆の傾向を示す．このため, SVM とこれらの手法との差は, C の値が小さくなるほど拡大する．従って, 両者を組み合わせた手法（seq 以外）を適用する場合は, あらかじめ C の最適値を決定しておく必要がある．JGSS-2000 データと JGSS-2001 データによるチューニングを行った結果, SVM は C=0.6, add-code は C=0.4, add-rule は C=0.3 となり, このときの正解率は, 図 1 においても評価できる値であった．

次に, seq についての調査結果を示して考察する．前述したように, 結果を利用する SVM の訓練データとしては, 全事例を用いる場合と決定できない事例のみを用いる場合の 2 通りがあるが, 表 1 から明らかなように, 全事例を用いる方が正解率が高い[8]．また, 利用される SVM の結果における正誤状況は, 表 2 に示すように, 全体では正しいものより間違う方がやや多く, 特に, ルールベース手法が未決定の場合に間違う場合が多い．ただし, ルールベース手法が複数個出力する場合は正しく決定する傾向がある．従って, seq において正解率を高くするには, 決定できない場合のケース別の処理を検討することが有効である．

表 1: seq で利用される事例による正解率の違い

	全事例	決定できない事例
C=1.0	72.9	71.0
max	73.1	71.1
min	72.9	70.5

表 2: seq で利用される SVM の正誤状況

ルールベース	SVM 正	SVM 誤	計
未決定コード	533	655	1188
複数個	328	255	583
計	861	910	1771

最後に, seq と add-code-rule 以外の手法において, 新しくコーディングした結果を訓練事例としてフィードバックする効果について述べる．表 3 より明らかなように, いずれの手法においても 1/20（約 340 サンプル）をフィードバックしたときに正解率の向上がみられ, その量が増えるほど効果的である．1/10 をフィードバックすると, いずれの手法においても約 1% 程度向上する．ここで, 実際のコーディング時に過去の事例が使えない場合もあるため, 単独のデータセットにより実験した結果, 訓練データと評価データが 1:9 のとき正解率は 56〜62% であるが, 1:1 のときに約 70〜74%, 2:1 のときに 72〜75% で, 10 分割交差検定の 9:1 のときの 75〜77% に近似した値であった．これらの値は, コーダにおける負担と正解率の観点から, コーディングすべき量の一つの目安を示したものといえる．

6 おわりに

本稿では, 職業コーディングに対して, SVM とルールベース手法の 4 種類の組み合わせによる適用を行った．その結果, 両者を組み合わせる方法はすべて, SVM を単独に適用するより正解率が高くなることがわかった．中でも, ルールベース手法が決定した職業コードを素性とする方法は最

[5] JGSS においては, 回答者 1 人に対して,「本人現職（無職の場合は最後職）」,「本人初職」,「配偶者職」の 3 種類の職業データを収集する．

[6] 調査年度の違いにより, 各データセットは性質が多少異なるが, 同質なデータが得られた状況も想定し, 単独のデータセット（JGSS-2002 データ）に対して 10 分割の交差検定による実験も行った．

[7] これは単独のデータセットにおいても同様である．

[8] この傾向は, 単独のデータセットにおいてさらに強かった．

表 3: フィードバックの効果（C=1.0 の場合）

	SVM	add-code	add-rule
ゼロ	71.9	73.1	72.8
1/10	72.9	74.2	74.0
1/11	72.7	74.0	73.6
1/15	72.4	73.8	73.5
1/20	72.3	73.7	73.3

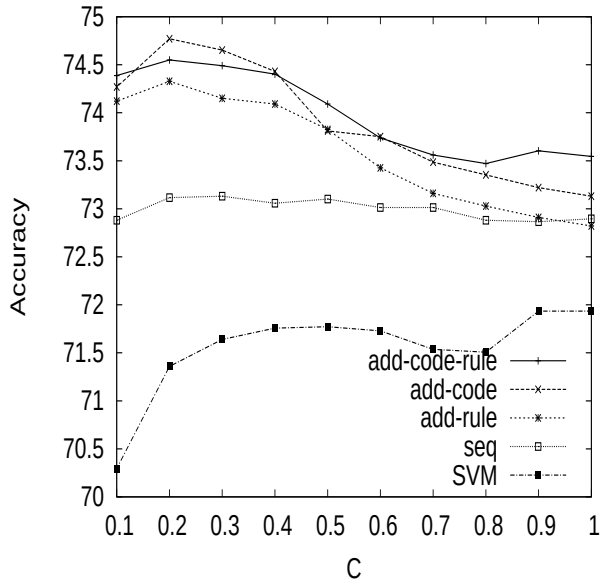

図 1: SVM とルールベース手法との組み合わせにおける正解率

も有効であった. また, 訓練データに過去の事例だけでなくコーディングが終了した事例を加えるフィードバックの方法が有効であることがわかった. 今後の課題は, SVM が出力する結果に対する確信度の付与についての問題を検討すること, また, フィードバックにおいて能動学習を取り入れることなどである.

謝辞

日本版 General Social Surveys（JGSS）は, 大阪商業大学比較地域研究所が, 文部科学省から学術フロンティア推進拠点としての指定を受けて（1999-2003 年度）, 東京大学社会科学研究所と共同で実施している研究プロジェクトである（研究代表：谷岡一郎・仁田道夫, 代表幹事：佐藤博樹・岩井紀子, 事務局長：大澤美苗）. データの入手先は, 東京大学社会科学研究所附属日本社会研究情報センター SSJ データ・アーカイブである.

データの利用にあたり, 関係者のご厚意に感謝いたします.

References

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. 1996. 1995 年 SSM 調査コード · ブック.

H. Isozaki and T. Hirao. 2003. Japanese Zero Pronoun Resolition based on Ranking Rules and Machine Learning. *Proceedings of the Eighth Conference on Empirical Methods in Natural Language Processing*, pp. 184–191.

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

工藤拓, 松本裕治. 2002. Support Vector Machine を用いた Chunk 同定. 自然言語処理　Vol.9 No.5, pp. 3–22.

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.

M. Núñez. 1991. The Use of Background Knowledge on Decision Tree Induction. *Machine Learning*, Vol.6, pp. 231–250.

大阪商業大学比較地域研究所, 東京大学社会科学研究所（編）. 2002. 日本版 General Social Surveys JGSS-2000 基礎集計表 · コードブック.

F. Sebastiani. 2002. Machine Learning Automated Text Categorization. *ACM Computing Surves*,Vol.34 No.1, pp. 1–47.

P. Seng-Bae and Z. Byoung-Tak. 2003. Text Chunking by Combining Hand-Crafted Rules and Memory-Based Learnung. *Proceedings of the 41-th Annual Meeting of the Association for Computational Linguistics(ACL 2003)*, pp. 497–504.

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－. 数理社会学会論文誌　理論と方法　Vol.15 No.1, pp. 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－. 敬愛大学国際研究　Vol.8, pp. 31–52.

高橋和子. 2002. 職業 · 産業コーディング自動化システムの活用. 第 8 回言語処理学会年次大会発表論文集, pp. 491–494.

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. 大阪商業大学比較地域研究所, 東京大学社会科学研究所（編）日本版 General Social Surves JGSS-2001 研究論文集［2］JGSS で見た日本人の意識と行動, pp. 179–191.

高橋和子, 高村大也, 奥村学. 2004. 機械学習とルールベースによる職業コーディング. 情報処理学会研究報告　2004-NL-159, pp. 53–60.

D. H. Wolpert. 1992. Stacked Generalization. *Neural Networks*, Vol.5, pp. 241–259.