ISSN 1348－7345

製表技術参考資料2

事業所・企業統計調査産業分類自動格付法の研究

サポートベクターマシンによる産業分類自動格付の研究

及び

国内外における統計分類自動格付法の研究動向

N S T A C

Working Paper No. 2

平成 16 年 8 月

独立行政法人　統計センター

製表技術参考資料は，独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

本資料の内容は，原則として職員が個人として執筆しており，機関の見解を示すものではない。

目　　次

Ⅰ． 事業所・企業統計調査産業分類自動格付法の研究

鈴木 清美*，髙崎 清*，岡本 政人*
磯部 祥子*，小野寺 夏代*，亀本 薫*，齋藤 なおみ*

要 旨

本稿は，統計センターが行う平成 16 年事業所・企業統計調査産業分類審査事務への自動格付法の利用可能性を検討するため，新産業分類符号に組替え済みの 13 年調査データを用いて実験を行った結果をまとめたものである。

主として地方における分類格付と現行自動格付システム[1]による分類格付が一致した場合に，当該分類符号を妥当と判定してよいかについて検証を行った。その結果，事業所産業分類については，自動格付システムが出力する第１候補及び第２候補の推定確率を基準に用いるよりも，幾つかの小分類区分に格付されしかも事業所の名称，事業の種類，取扱商品に特定キーワードが含まれる事業所と製造業に格付された事業所を除外したほうが地方格付と一致する割合（一致率）及び一致した場合の正解率が高くなり，正解率 99%以上を許容範囲とすると，自動格付システムで 60%以上カバーできることが明らかとなった。

さらに，事業の種類＋取扱商品（従来方式）で学習・格付を行った結果と，事業の種類のみで学習・格付を行った結果のうち，後者で全文一致方式の格付ができた場合これを用い，他の場合はそれぞれの第１候補の推定確率が高いほうを用いる合成方式を採用することで，正解率を落とさずに一致率を２ポイント程度向上させることができた。

企業産業分類については，地方における分類格付と自動格付が一致した場合，推定確率や産業区分・特定キーワードなどによる除外を行わずに高い正解率が確保でき，99%以上を許容範囲とすると，自動格付システムで 60%近くをカバーできることが明らかとなった。

さらに，企業全体の事業の種類（従来方式）で学習・格付を行った結果と，事業所としての事業の種類＋取扱商品で学習・格付を行った結果のうち，前者で全文一致方式の格付ができた場合これを用い，他の場合はそれぞれの第１候補の推定確率が高いほうを用いる合成方式を採用すると，正解率 99%以上を維持しつつ一致率を 66%まで高めることができ

[1] 現行の産業分類自動格付システムについては，戸井田&瀬谷[1996]，米澤[2000]参照。

た。

自動格付の導入効果をより高めるには，分類性能をさらに向上させる必要がある。現在知られている自動格付の方法論で分類性能を飛躍的に向上させることは困難と思われるが，今回の実験の過程で幾つか工夫の余地があることも判明したことから，本稿にまとめた結果を踏まえ，今後さらに研究を進めていく。

* 統計センター研究センター
E-mail: research@nstac.go.jp

Ⅰ． 事業所・企業統計調査産業分類自動格付法の研究

鈴木 清美，髙崎 清，岡本 政人
磯部 祥子，小野寺 夏代，亀本 薫，齋藤 なおみ

1 実験・検証目的

統計センターが行う平成 16 年事業所・企業統計調査産業分類審査への自動格付システムの利用の可能性を検討した。

2 実験対象等

- ２県を対象に、平成 13 年事業所・企業統計調査の県から提出の個票データを用いて行った。
- 格付対象は、16 年調査の符号格付予定範囲に合わせて、新設事業所及び、継続事業所のうち「事業の種類」（産業分類符号）が平成 11 年調査から変わった変動事業所とした。なお、事業所産業分類については、商業統計調査対象の卸売・小売業を除いた。
- 産業分類符号は、新分類符号（平成 14 年３月日本標準産業分類改訂後の符号）に組替えたものを使用した。

3 格付ルール生成

実験対象の２県のほか、この両県に近・隣接する県を加えた５県のデータを用いてルール生成を行った。

なお、実験対象の２県からは、格付対象事業所データを除外した。

4 前処理

表記のゆれ（例えば「らーめん」と「ラーメン」）をなるべく減らすため、自動格付用及び格付ルール生成用データの「事業の種類」及び「取扱商品」に対して表記を統一する前処理を行った。

5 検証結果の概要 （従来方式による格付結果）

主として、自動格付と地方格付が一致した場合に、地方格付が適切であると判断してよいか、そして、一致する割合がどの程度か、という観点から分析を行った。

1）事業所産業

表Ⅰ－1 自動格付と地方格付の一致率及び正解率（事業所産業） （%）

県名	推薦基準値等	合 計		新設事業所		変動事業所	
		一致率	正解率	一致率	正解率	一致率	正解率
A県	推薦基準値指定なし	72.34	98.43	79.66	98.18	60.48	98.97
	第１位 0.8 以上、第２位 0.3 未満	53.71	98.87	63.63	98.79	37.64	99.09
	第１位 0.97 以上、第２位 0.3 未満	42.78	99.11	52.93	99.09	26.34	99.17
	推薦基準値指定なし（一部産業除外）	55.44	99.35	63.23	99.39	42.83	99.27
	推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	61.94	99.50	71.48	99.49	46.49	99.53
B県	推薦基準値指定なし	68.95	98.64	74.06	98.63	56.68	98.68
	第１位 0.8 以上、第２位 0.3 未満	51.78	99.30	58.72	99.31	35.09	99.27
	第１位 0.97 以上、第２位 0.3 未満	41.63	99.49	48.64	99.54	24.79	99.29
	推薦基準値指定なし（一部産業除外）	53.17	99.33	58.67	99.39	39.96	99.14
	推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	60.98	99.40	67.60	99.43	45.08	99.29

［注］ 一致率は、格付対象事業所（総数）に占める自動格付と地方格付が一致した事業所の比率であり、正解率は、その一致事業所のうち、正解（人手審査格付と一致したもの）の比率である。

なお、「推薦基準値指定なし（一部産業除外）」及び「推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）」の一致率計算の母数となる格付対象事業所数についても、除外した一部特定産業又はｷｰﾜｰﾄﾞに該当する事業所を含めている。

ア 自動格付と地方格付の一致率及び正解率

⑺ （単語分割方式の格付結果について）推薦基準値指定なし(別表Ⅰ-1-1)の場合でみると、自動格付と地方格付との一致率［注］はA県 72.34%、B県 68.95%となっている。一致の場合の正解率(人手審査格付との一致率)は、A県 98.43%、B県 98.64%である。

新設・変動事業所別にみると、一致率は、２県とも新設が 70%台（それぞれ 79.66%、74.06%）、変動が 60%程度（同 60.48%、56.68%）と変動が 17～19 ポイント低い。正解率は、新設がA県 98.18%、B県 98.63%、変動がそれぞれ 98.97%、98.68%と、いずれも 98%台となっている。

［注］格付対象事業所数に対する自動格付と地方格付との一致数の比率をいう。以下同じ。

(ｲ)　推薦基準値第１位 0.8 以上・第２位 0.3 未満（以下「基準値 0.8・0.3」という。）とすると、一致率はＡ県 53.71%、Ｂ県 51.78%と 50%台となる。新設・変動別にみると、新設がＡ県 63.63%、Ｂ県 58.72%に対し、変動がそれぞれ 37.64%、35.09%と新設よりも 25 ポイント程度低い。

正解率は、Ａ県 98.87%、Ｂ県 99.30%となり、推薦基準値指定なしの場合より高くなっている。新設･変動別にみると、新設がＡ県 98.79%、Ｂ県 99.31%、変動がそれぞれ 99.09%、99.27%と、Ａ県の新設以外は 99%台と高率を示す。

(ｳ)　推薦基準値を更に厳しくして 0.97・0.3(別表Ⅰ-1-2)とすると、一致率は２県ともそれぞれ 42.78%、41.63%と 40%台に下がる。一方、正解率はそれぞれ 99.11%、99.49%となり、基準値 0.8・0.3 の場合に比べ、やや高くなる。このように、推薦基準値を厳しくすると、正解率は向上するが、一致率の低下幅が大きい。

イ 自動格付と地方格付が一致しない割合（不一致率）及び不一致の場合の自動格付の正解率

推薦基準値指定なし（別表Ⅰ-2）で、自動格付と地方格付が一致しない割合（不一致率）は、Ａ県 20.50%、Ｂ県 23.73%、不一致の場合の自動格付の正解率はそれぞれ 7.85%、12.56%と、正解率が極端に低い。新設・変動事業所別にみても、大差ない。また、基準値 0.8・0.3 とすると、不一致率は 4～6%、正解率は 20～30%となり、更に基準値 0.97・0.3 とすると、不一致率は 2～4%、正解率は 25～35%となり、いずれも正解率は低率である。

このように、自動格付の正解率が低率のため、不一致のものはすべて人手審査が必要とみられる。

［注］企業産業についても同様なため、自動格付と地方格付が一致しない場合の記述は省略。

ウ 産業大分類別の一致率及び正解率

推薦基準値指定なしの場合で、産業大分類別に自動格付と地方格付との一致率及び一致の場合の正解率をみると、製造業の一致率、正解率とも他産業と比べ総じて低い。

表Ⅰ－2 製造業の自動格付と地方格付との一致率及び正解率（推薦基準値指定なし）

		新設事業所					変動事業所				
		格付対象数	格付一致数	一致率(%)	うち正解数	正解率(%)	格付対象数	格付一致数	一致率(%)	うち正解数	正解率(%)
A県	全産業	5,186	4,131	79.66	4,056	98.18	3,201	1,936	60.48	1,916	98.97
	製造業	453	238	**52.54**	224	**94.12**	747	340	**45.52**	333	**97.94**
B県	全産業	30,149	22,329	74.06	22,022	98.63	12,548	7,112	56.68	7,018	98.68
	製造業	2,570	1,225	**47.67**	1,172	**95.67**	2,533	1,026	**40.51**	1,000	**97.47**

エ 一部産業を除外することによる格付精度の改善

自動格付と地方格付は一致しているが、人手審査により不正解（誤り）と判定されたものが比較的多い産業小分類（自動格付及び地方格付が付与した分類区分、別表１参照）を自動格付対象から除外し、集計した結果は第４表のとおりである。

除外した産業以外の平均は、推薦基準値指定なしの場合で、一致率がＡ県 55.44%、Ｂ県 53.17%となり、除外前（それぞれ 72.34%、68.95%）よりは低くなるが、正解率は、Ａ県 99.35%、Ｂ県 99.33%となり、アの(ｲ)、(ｳ)で述べた推薦基準値を設定した場合よりも一致率は高く、正解率は同程度となっている。

新設・変動事業所別にみると、一致率は新設がＡ県 63.23%、Ｂ県 58.67%、変動がそれぞれ 42.83%、39.96%と、除外前と同様にいずれも変動が 19～20 ポイント低い。正解率は、新設がＡ県、Ｂ県とも 99.39%、変動がそれぞれ 99.27%、99.14%と、いずれも 99%を上回っている。

オ 特定キーワードを除外することによる格付精度の改善

格付の誤りやすい小分類で、人手審査により不正解（誤り）と判定されたものの中には、誤りと判断される根拠となる特定キーワード（別表Ⅰ-7 参照）を含んでいるものが多いことから、産業小分類（自動格付及び地方格付が付与した分類区分）ごとに、特定キーワードを含む事業所のみを除外し、集計した結果は別表Ⅰ-5 のとおりである。

なお、キーワードを特定できない小分類は、小分類単位（製造業は大分類単位）で除外した。

除外した事業所以外の平均は、推薦基準値指定なしの場合で、一致率がＡ県 61.94%、Ｂ県 60.98%となり、小分類単位で除外した場合（それぞれ 55.44%、53.17%）よりも７ポイント前後高くなる。また、正解率も、Ａ県 99.50%、Ｂ県 99.40%となり、小分類単位で除外した場合（それぞれ 99.35%、99.33%）を 0.1 ポイント程度上回っている。

2）企業産業

表Ⅰ－3 自動格付と地方格付の一致率及び正解率（企業産業） （%）

県名	推薦基準値等	合 計		新設事業所		変動事業所	
		一致率	正解率	一致率	正解率	一致率	正解率
Ａ県	推薦基準値指定なし	62.33	99.68	57.22	98.06	63.43	100.00
	第１位 0.8 以上、第２位 0.3 未満	42.70	99.77	43.33	98.72	42.57	100.00
	第１位 0.97 以上、第２位 0.3 未満	35.11	100.00	37.78	100.00	34.53	100.00
	推薦基準値指定なし（一部産業除外）	58.48	99.66	53.89	97.94	59.47	100.00
	推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	61.44	99.84	56.67	99.02	62.47	100.00
Ｂ県	推薦基準値指定なし	58.12	99.27	55.01	97.70	59.12	99.73
	第１位 0.8 以上、第２位 0.3 未満	41.24	99.77	39.05	99.25	41.94	99.93
	第１位 0.97 以上、第２位 0.3 未満	34.15	99.93	30.77	99.68	35.24	100.00
	推薦基準値指定なし（一部産業除外）	54.14	99.47	50.24	98.64	55.39	99.72
	推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	56.51	99.45	53.16	98.53	57.59	99.73

［注］ 一致率は、格付対象事業所（総数）に占める自動格付と地方格付が一致した事業所の比率であり、正解率は、その一致事業所のうち、正解（人手審査格付と一致したもの）の比率である。

なお、「推薦基準値指定なし（一部産業除外）」及び「推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）」の一致率計算の母数となる格付対象事業所数についても、除外した一部特定産業又はｷｰﾜｰﾄﾞに該当する事業所を含めている。

ア 自動格付と地方格付の一致率及び正解率

(ｱ) （単語分割方式の格付結果について）推薦基準値指定なし(別表Ⅰ-1-1)の場合でみると、自動格付と地方格付との一致率は、Ａ県 62.33%、Ｂ県 58.12%となっている。一致の場合の正解率は、Ａ県 99.68%、Ｂ県 99.27%と高率である。新設・変動事業所別にみると、一致率、正解率とも新設が変動を下回っている。

(ｲ) 基準値 0.8・0.3 とすると、一致率はＡ県 42.70%、Ｂ県 41.24%と４割程度に

下がるが、正解率は２県とも 99.77%と 100%近くに上昇する。

(ｳ) 基準値を更に厳しくして 0.97･0.3(別表Ⅰ-1-2)とすると、一致率はＡ県 35.11%、Ｂ県 34.15%と３割半ばに下がるが、正解率はＡ県 100%、Ｂ県 99.93%となり、ほぼ完璧な正解率となっている。これは、全文一致方式の格付結果が大部分を占めるようになるためである。

イ 産業別の一致率及び正解率と、一部産業、特定キーワードを除外した場合の格付精度

(ｱ) 推薦基準値指定なしの場合、自動格付と地方格付の一致率は、事業所産業の場合と同様に、製造業が 30～40%程度と全産業平均より 20～30 ポイント程度低いが、正解率は 99～100%と高い。

(ｲ) 事業所産業の場合と同様に、特定の産業（事業所産業分類の場合に除外したものと同じ産業、ただし「Ｆ製造業」は含めた。）（別表Ⅰ-7）を自動格付対象から除外し、集計した結果は第４表のとおりである。

除外した産業以外の平均は、推薦基準値指定なしの場合で、一致率がＡ県 58.48%、Ｂ県 54.14%となり、除外前（それぞれ 62.33%、58.12%）よりそれぞれ４ポイント程度低くなる。その一方で、正解率はＡ県 99.66%、Ｂ県 99.47%となり、特に目立った効果はみられない。

(ｳ) 事業所産業の場合と同様に、産業小分類ごとに特定キーワード（事業所産業分類の場合と同じキーワード、ただし「Ｆ製造業」は含めた。）（別表Ⅰ-8）を含む事業所のみを除外し、集計した結果は第５表のとおりである。

除外した事業所以外の平均は、推薦基準値指定なしの場合で、一致率がＡ県 61.44%、Ｂ県 56.51%となり、小分類単位で除外した場合よりも２～３ポイント高くなる。正解率は、Ａ県 99.84%、Ｂ県 99.45%となり、小分類単位で除外した場合とほとんど変わらないが、全く除外しない場合と比べると 0.2 ポイント程度高くなっている。

６　検証結果の概要 （事業の種類のみと、事業の種類 ＋ 取扱商品の格付結果を組合せた場合）

1）事業所産業

事業所産業について、取扱商品を含めた場合(従来方式)と、事業の種類のみの２種類の自動格付を行い、事業の種類のみの場合に全文一致方式で格付可能な事業所についてはその格付結果を、全文一致方式で格付不可能な事業所については、単語分割方式格付結果の第１位候補の推薦確率が従来方式の第１位候補の推薦確率よりも高い場合、その格付結果を採用する合成方式を適用した結果を、表Ｉ－４に示す。

推薦基準値指定なしの場合の一致率は、A県 76.77%、B県 71.13%となり、従来方式（それぞれ 72.34%、68.95%）を 2～4.5 ポイント上回っている。正解率はA県 98.34%、B県 98.64%で、従来方式（それぞれ 98.43%、98.64%）と同程度である。

産業小分類ごとに特定キーワードを含む事業所を除外した場合の一致率は、A県 65.27%、B県 62.53%となり、従来方式（それぞれ 61.94%、60.98%）を 1.5～3.5 ポイント上回っている。また、正解率はA県 99.36%、B県 99.43%となり、従来方式（それぞれ 99.50%、99.40%）と同程度である。

表Ｉ－4　合成方式と従来方式の比較 (事業所産業)　　（%）

県名	推薦基準値等		合 計		新設事業所		変動事業所	
			一致率	正解率	一致率	正解率	一致率	正解率
A県	従来方式	推薦基準値指定なし	72.34	98.43	79.66	98.18	60.48	98.97
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	61.94	99.50	71.48	99.49	46.49	99.53
	合成方式	推薦基準値指定なし	76.77	98.34	82.66	98.06	67.23	98.88
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	65.27	99.36	73.52	99.34	51.89	99.40
B県	従来方式	推薦基準値指定なし	68.95	98.64	74.06	98.63	56.68	98.68
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	60.98	99.40	67.60	99.43	45.08	99.29
	合成方式	推薦基準値指定なし	71.13	98.64	75.89	98.65	59.67	98.58
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	62.53	99.43	69.03	99.46	46.91	99.32

脚注は、表Ｉ－1 参照。

2）企業産業

企業産業について、会社全体の事業の種類のみの場合(従来方式)と、事業所情報（事業の種類及び取扱商品）を用いた場合の２種類の自動格付を行い、従来方式で全文一致方式の格付不可能な事業所について、後者による自動格付の第１位候補の推薦確率が従来方式の第１位候補の推薦確率よりも高い場合、その格付結果を採用する合成方式を適用した結果を、表Ｉ－５に示す。

推薦基準値指定なしの場合の一致率は、A県 71.30%、B県 65.14%と、従来方式（それぞれ 62.33%、58.12%）を 7～9 ポイント上回っており、事業所産業の場合よりも上昇幅が大きくなっている。一方、正解率はA県 99.45%、B県 99.13%となり、従来方式（それぞれ 99.68%、99.27%）をやや下回っている。

産業小分類ごとに特定キーワードを含む事業所を除外した場合の一致率は、A県 70.12%、B県 63.13%と、従来方式（それぞれ 61.44%、56.51%）を 6.5～8.5 ポイント上回っており、事業所産業の場合よりも上昇幅が大きくなっている。一方、正解率は、A県 99.58%、B県 99.25%となり、従来方式（それぞれ 99.84%、99.45%）をやや下回っている。

このように、合成方式では正解率がやや低下するが、99%を上回っている。

表Ｉ－5 合成方式と従来方式の比較（企業産業）　　（%）

県名	推薦基準値等		合 計		新設事業所		変動事業所	
			一致率	正解率	一致率	正解率	一致率	正解率
A県	従来方式	推薦基準値指定なし	62.33	99.68	57.22	98.06	63.43	100.00
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	61.44	99.84	56.67	99.02	62.47	100.00
	合成方式	推薦基準値指定なし	71.30	99.45	65.56	98.31	72.54	99.67
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	70.12	99.58	65.00	99.15	71.22	99.66
B県	従来方式	推薦基準値指定なし	58.12	99.27	55.01	97.70	59.12	99.73
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	56.51	99.45	53.16	98.53	57.59	99.73
	合成方式	推薦基準値指定なし	65.14	99.13	61.54	96.99	66.30	99.76
		推薦基準値指定なし（ｷｰﾜｰﾄﾞによる除外）	63.13	99.25	59.69	97.55	64.23	99.76

脚注は、表Ｉ－３参照。

7 結 論

統計センターが行う産業分類格付審査事務に自動格付システムを利用することの適否は、自動格付システムの判定結果に対する許容精度に依存する。正解率 99%以上を許容範囲と仮定すると、企業産業分類の場合、格付審査対象事業所全体の 60%程度を、さらに、合成方式により 65%程度を自動格付システムで処理できると期待される。事業所産業分類についても、誤りと判断される根拠となる特定キーワードを含む事業所を除外することにより、全体の 60%程度を、合成方式では 62%程度を自動格付システムで処理できると期待される。

自動格付でカバーできる範囲を広げるには分類性能をさらに向上させる必要がある。現在知られている自動格付の方法論で分類性能を飛躍的に向上させることは困難と思われるが，今回の実験の過程で幾つか工夫の余地があることも判明したことから，本稿にまとめた結果を踏まえ，今後さらに研究を進めていく。

なお、正解率は高いほどよいが、推薦基準値を厳しくしても正解率をさらに引き上げる効果は小さい。正解率 100%が求められる場合は、全文一致方式による自動格付に限定するしかない。全文一致方式で自動格付できるのは、事業所産業の場合で格付審査対象事業所全体の２割程度、企業産業の場合で３割弱であるが、事業所産業分類について「事業の種類」だけの全文一致（さらに、「取扱い商品」も含めた単語分割方式の格付結果と一致する場合に限定する方式も考えられる）とした場合、100%近い正解率が得られ、35%程度カバーできる可能性がある。

したがって、正解率をより重視するのであれば、（ある種の）全文一致方式に限定することは、一つの選択肢になろう。

参 考 文 献

戸井田 幸記, 瀬谷 恵子[1996]. 産業分類の自動格付技法に関する研究, 統計局研究彙報, 第 54 号，pp.87-136.

米澤 哲一[2000]. 産業分類自動格付システムの７年間（平成４～１０年度）の研究について，統計局研究彙報，第 59 号，pp.61-97.

別 表

別表Ｉ－１－１　自動格付と地方格付との一致数（率）及びその正解数（率）

（推薦基準値指定なし）

格付対象事業所数				自動格付数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
								自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
A県	事業所産業	合計	8,387	7,786	6,107	6,067	5,972	1,877	1,873	4,190	4,099
			100.00	92.83	78.44	72.34	98.43	22.38	99.79	49.96	97.83
		新設	5,186	4,911	4,152	4,131	4,056	1,507	1,504	2,624	2,552
			100.00	94.70	84.54	79.66	98.18	29.06	99.80	50.60	97.26
		変動	3,201	2,875	1,955	1,936	1,916	370	369	1,566	1,547
			100.00	89.82	68.00	60.48	98.97	11.56	99.73	48.92	98.79
	企業産業	合計	1,014	881	651	632	630	307	307	325	323
			100.00	86.88	73.89	62.33	99.68	30.28	100.00	32.05	99.38
		新設	180	159	114	103	101	61	61	42	40
			100.00	88.33	71.70	57.22	98.06	33.89	100.00	23.33	95.24
		変動	834	722	537	529	529	246	246	283	283
			100.00	86.57	74.38	63.43	100.00	29.50	100.00	33.93	100.00
B県	事業所産業	合計	42,697	39,574	30,313	29,441	29,040	9,970	9,952	19,471	19,088
			100.00	92.69	76.60	68.95	98.64	23.35	99.82	45.60	98.03
		新設	30,149	28,280	23,049	22,329	22,022	8,494	8,481	13,835	13,541
			100.00	93.80	81.50	74.06	98.63	28.17	99.85	45.89	97.87
		変動	12,548	11,294	7,264	7,112	7,018	1,476	1,471	5,636	5,547
			100.00	90.01	64.32	56.68	98.68	11.76	99.66	44.92	98.42
	企業産業	合計	4,217	3,594	2,521	2,451	2,433	1,232	1,231	1,219	1,202
			100.00	85.23	70.14	58.12	99.27	29.22	99.92	28.91	98.61
		新設	1,027	877	600	565	552	270	269	295	283
			100.00	85.39	68.42	55.01	97.70	26.29	99.63	28.72	95.93
		変動	3,190	2,717	1,921	1,886	1,881	962	962	924	919
			100.00	85.17	70.70	59.12	99.73	30.16	100.00	28.97	99.46

＊ 事業所産業用ルールは、商業統計調査対象の卸売・小売業を含まないルール

（注）１　「格付対象事業所数」は、新設・変動事業所（ただし、事業所産業の場合は商業統計調査対象の卸売・小売業を除く）の総数。

２　「自動格付数」は、格付対象事業所のうち、自動格付されなかったものを除いた数。

３　「自動格付と地方格付との一致数」は、自動格付されたもののうち、自動格付と地方格付が一致した数。ここでいう一致率は、格付対象事業所数に対する一致数の比率。

４　「人手審査格付との一致数」は、「自動格付数」又は「自動格付と地方格付との一致数」のうちの「人手審査格付との一致数」であり、いわゆる「正解数」のこと。

また、ここでいう「人手審査格付との一致率」は、いわゆる「正解率」のこと。**（Ｉ－１－２表に続く）**

別表Ｉ－１－２　自動格付と地方格付との一致数（率）及びその正解数（率）
（推薦基準値第 1 位 0.97 以上，第 2 位 0.3 未満）

格付対象事業所数				自動格付数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
								自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
A県	事業所産業	合計	8,387	3,754	3,602	3,588	3,556	1,877	1,873	1,711	1,683
			100.00	44.76	95.95	42.78	99.11	22.38	99.79	20.40	98.36
		新設	5,186	2,836	2,757	2,745	2,720	1,507	1,504	1238	1216
			100.00	54.69	97.21	52.93	99.09	29.06	99.80	23.87	98.22
		変動	3,201	918	845	843	836	370	369	473	467
			100.00	28.68	92.05	26.34	99.17	11.56	99.73	14.78	98.73
	企業産業	合計	1,014	395	363	356	356	307	307	49	49
			100.00	38.95	91.90	35.11	100.00	30.28	100.00	4.83	100.00
		新設	180	81	73	68	68	61	61	7	7
			100.00	45.00	90.12	37.78	100.00	33.89	100.00	3.89	100.00
		変動	834	314	290	288	288	246	246	42	42
			100.00	37.65	92.36	34.53	100.00	29.50	100.00	5.04	100.00
B県	事業所産業	合計	42,697	19,266	18,182	17,776	17,686	9,970	9,952	7806	7734
			100.00	45.12	94.37	41.63	99.49	23.35	99.82	18.28	99.08
		新設	30,149	15,541	15,016	14,665	14,597	8,494	8,481	6171	6116
			100.00	51.55	96.62	48.64	99.54	28.17	99.85	20.47	99.11
		変動	12,548	3,725	3,166	3,111	3,089	1,476	1,471	1635	1618
			100.00	29.69	84.99	24.79	99.29	11.76	99.66	13.03	98.96
	企業産業	合計	4,217	1,601	1,473	1,440	1,439	1,232	1,231	208	208
			100.00	37.97	92.00	34.15	99.93	29.22	99.92	4.93	100.00
		新設	1,027	357	331	316	315	270	269	46	46
			100.00	34.76	92.72	30.77	99.68	26.29	99.63	4.48	100.00
		変動	3,190	1,244	1,142	1,124	1,124	962	962	162	162
			100.00	39.00	91.80	35.24	100.00	30.16	100.00	5.08	100.00

（Ｉ－１－１表の脚注の続き）

５　率の単位は%

別表Ｉ－２　自動格付と地方格付が不一致の場合の自動格付正解数（率）

（推薦基準値指定なし）

格付対象事業所数				自動格付と地方格付との不一致数（率）	自動格付と人手審査格付との一致数（率）	全文一致方式		単語分割方式	
						自動格付と地方格付との不一致数（率）	自動格付と人手審査格付との一致数（率）	自動格付と地方格付との不一致数（率）	自動格付と人手審査格付との一致数（率）
A県	事業所産業	合計	8,387	1,719	135	64	25	1,655	110
			100.00	20.50	7.85	0.76	39.06	19.73	6.65
		新設	5,186	780	96	40	18	740	78
			100.00	15.04	12.31	0.77	45.00	14.27	10.54
		変動	3,201	939	39	24	7	915	32
			100.00	29.33	4.15	0.75	29.17	28.58	3.50
	企業産業	合計	1,014	249	21	31	7	218	14
			100.00	24.56	8.43	3.06	22.58	21.50	6.42
		新設	180	56	13	10	5	46	8
			100.00	31.11	23.21	5.56	50.00	25.56	17.39
		変動	834	193	8	21	2	172	6
			100.00	23.14	4.15	2.52	9.52	20.62	3.49
B県	事業所産業	合計	42,697	10,133	1,273	730	250	9,403	1,023
			100.00	23.73	12.56	1.71	34.25	22.02	10.88
		新設	30,149	5,951	1,027	438	213	5,513	814
			100.00	19.74	17.26	1.45	48.63	18.29	14.77
		変動	12,548	4,182	246	292	37	3,890	209
			100.00	33.33	5.88	2.33	12.67	31.00	5.37
	企業産業	合計	4,217	1,143	88	117	25	1,026	63
			100.00	27.10	7.70	2.77	21.37	24.33	6.14
		新設	1,027	312	48	31	13	281	35
			100.00	30.38	15.38	3.02	41.94	27.36	12.46
		変動	3,190	831	40	86	12	745	28
			100.00	26.05	4.81	2.70	13.95	23.35	3.76

＊ 事業所産業用ルールは、商業統計調査対象の卸売・小売業を含まないルール

（注）１　「格付対象事業所数」は、新設・変動事業所（ただし、事業所産業の場合は商業統計調査対象の卸売・小売業を除く）の総数。

２　「自動格付と地方格付との不一致数」は、自動格付されたもののうち、自動格付と地方格付との不一致数。ここでいう不一致率は、格付対象事業所数に対する不一致数の比率。

３　「自動格付と人手審査格付との一致数」は、「自動格付と地方格付との不一致数」のうちの「自動格付と人手審査格付との一致数」であり、いわゆる「自動格付の正解数」のこと。
また、ここでいう「自動格付と人手審査格付との一致率」は、いわゆる「自動格付の正解率」のこと。

４　率の単位は%

別表Ｉ－３－１　事業所産業大分類別自動格付と地方格付との一致数（率）及びその正解数（率）

A県新設事業所産業　　（推薦基準値指定なし）

格付対象事業所数		自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
				自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
合　　計	5,186	4,131	4,056	1,507	1,504	2,624	2,552
	100.00	79.66	98.18	29.06	99.80	50.60	97.26
A 農業	26	19	19	2	2	17	17
	100.00	73.08	100.00	7.69	100.00	65.38	100.00
B 林業	0	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
C 漁業	3	3	3	1	1	2	2
	100.00	100.00	100.00	33.33	100.00	66.67	100.00
D 鉱業	3	2	2	0	0	2	2
	100.00	66.67	100.00	0.00	0.00	66.67	100.00
E 建設業	761	578	568	162	162	416	406
	100.00	75.95	98.27	21.29	100.00	54.66	97.60
F 製造業	453	238	224	15	15	223	209
	100.00	52.54	94.12	3.31	100.00	49.23	93.72
G 電気・ガス・熱供給・水道業	7	2	1	0	0	2	1
	100.00	28.57	50.00	0.00	0.00	28.57	50.00
H 情報通信	156	122	122	37	37	85	85
	100.00	78.21	100.00	23.72	100.00	54.49	100.00
I 運輸業	171	118	117	20	20	98	97
	100.00	69.01	99.15	11.70	100.00	57.31	98.98
J 卸売・小売業	46	26	26	5	5	21	21
	100.00	56.52	100.00	10.87	100.00	45.65	100.00
K 金融・保険業	150	129	129	58	58	71	71
	100.00	86.00	100.00	38.67	100.00	47.33	100.00
L 不動産業	287	230	228	93	93	137	135
	100.00	80.14	99.13	32.40	100.00	47.74	98.54
M 飲食店，宿泊業	1,148	1,023	1,019	387	387	636	632
	100.00	89.11	99.61	33.71	100.00	55.40	99.37
N 医療，福祉	355	321	305	133	133	188	172
	100.00	90.42	95.02	37.46	100.00	52.96	91.49
O 教育，学習支援業	284	255	253	155	155	100	98
	100.00	89.79	99.22	54.58	100.00	35.21	98.00
P 複合サービス業	95	89	88	54	54	35	34
	100.00	93.68	98.88	56.84	100.00	36.84	97.14
Q サービス業（他に分類されないもの）	1,241	976	952	385	382	591	570
	100.00	78.65	97.54	31.02	99.22	47.62	96.45

別表Ｉ－１－１の脚注参照

別表Ｉ－３－２　事業所産業大分類別自動格付と地方格付との一致数（率）及びその正解数（率）

A県変動事業所産業　　（推薦基準値指定なし）

格付対象事業所数		自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
				自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
合　　計	3,201	1,936	1,916	370	369	1,566	1,547
	100.00	60.48	98.97	11.56	99.73	48.92	98.79
A 農業	9	5	5	0	0	5	5
	100.00	55.56	100.00	0.00	0.00	55.56	100.00
B 林業	1	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
C 漁業	0	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
D 鉱業	8	6	6	0	0	6	6
	100.00	75.00	100.00	0.00	0.00	75.00	100.00
E 建設業	741	481	480	125	125	356	355
	100.00	64.91	99.79	16.87	100.00	48.04	99.72
F 製造業	747	340	333	23	23	317	310
	100.00	45.52	97.94	3.08	100.00	42.44	97.79
G 電気・ガス・熱供給・水道業	5	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
H 情報通信	67	49	49	9	9	40	40
	100.00	73.13	100.00	13.43	100.00	59.70	100.00
I 運輸業	81	44	44	5	5	39	39
	100.00	54.32	100.00	6.17	100.00	48.15	100.00
J 卸売・小売業	26	12	12	5	5	7	7
	100.00	46.15	100.00	19.23	100.00	26.92	100.00
K 金融・保険業	38	27	27	6	6	21	21
	100.00	71.05	100.00	15.79	100.00	55.26	100.00
L 不動産業	190	126	126	32	32	94	94
	100.00	66.32	100.00	16.84	100.00	49.47	100.00
M 飲食店，宿泊業	458	362	360	86	86	276	274
	100.00	79.04	99.45	18.78	100.00	60.26	99.28
N 医療，福祉	34	27	27	7	7	20	20
	100.00	79.41	100.00	20.59	100.00	58.82	100.00
O 教育，学習支援業	51	35	34	17	17	18	17
	100.00	68.63	97.14	33.33	100.00	35.29	94.44
P 複合サービス業	35	22	20	7	6	15	14
	100.00	62.86	90.91	20.00	85.71	42.86	93.33
Q サービス業（他に分類されないもの）	710	400	393	48	48	352	345
	100.00	56.34	98.25	6.76	100.00	49.58	98.01

別表Ｉ－１－１の脚注参照

別表Ｉ－３－３　事業所産業大分類別自動格付と地方格付との一致数（率）及びその正解数（率）

B県新設事業所産業　　（推薦基準値指定なし）

格付対象事業所数		自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
				自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
合　　計	30,149	22,329	22,022	8,494	8,481	13,835	13,541
	100.00	74.06	98.63	28.17	99.85	45.89	97.87
A 農業	63	22	22	1	1	21	21
	100.00	34.92	100.00	1.59	100.00	33.33	100.00
B 林業	4	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
C 漁業	2	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
D 鉱業	5	2	2	0	0	2	2
	100.00	40.00	100.00	0.00	0.00	40.00	100.00
E 建設業	3,459	2,117	2,098	452	451	1,665	1,647
	100.00	61.20	99.10	13.07	99.78	48.14	98.92
F 製造業	2,570	1,225	1,172	123	123	1,102	1,049
	100.00	47.67	95.67	4.79	100.00	42.88	95.19
G 電気・ガス・熱供給・水道業	12	2	2	2	2	0	0
	100.00	16.67	100.00	16.67	100.00	0.00	0.00
H 情報通信	800	582	576	235	234	347	342
	100.00	72.75	98.97	29.38	99.57	43.38	98.56
I 運輸業	816	459	456	88	88	371	368
	100.00	56.25	99.35	10.78	100.00	45.47	99.19
J 卸売・小売業	201	28	25	5	5	23	20
	100.00	13.93	89.29	2.49	100.00	11.44	86.96
K 金融・保険業	810	659	654	295	294	364	360
	100.00	81.36	99.24	36.42	99.66	44.94	98.90
L 不動産業	2,395	1,765	1,753	735	734	1,030	1,019
	100.00	73.70	99.32	30.69	99.86	43.01	98.93
M 飲食店，宿泊業	8,348	7,131	7,069	3,075	3,074	4,056	3,995
	100.00	85.42	99.13	36.84	99.97	48.59	98.50
N 医療，福祉	1,928	1,684	1,617	713	712	971	905
	100.00	87.34	96.02	36.98	99.86	50.36	93.20
O 教育，学習支援業	1,788	1,492	1,468	714	713	778	755
	100.00	83.45	98.39	39.93	99.86	43.51	97.04
P 複合サービス業	369	326	323	85	84	241	239
	100.00	88.35	99.08	23.04	98.82	65.31	99.17
Q サービス業（他に分類されないもの）	6,579	4,835	4,785	1,971	1,966	2,864	2,819
	100.00	73.49	98.97	29.96	99.75	43.53	98.43

別表Ｉ－１－１の脚注参照

別表Ｉ－３－４　事業所産業大分類別自動格付と地方格付との一致数（率）及びその正解数（率）

Ｂ県変動事業所産業　　（推薦基準値指定なし）

格付対象事業所数		自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
				自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
合　　計	12,548	7,112	7,018	1,476	1,471	5,636	5,547
	100.00	56.68	98.68	11.76	99.66	44.92	98.42
A 農業	20	8	8	2	2	6	6
	100.00	40.00	100.00	10.00	100.00	30.00	100.00
B 林業	1	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
C 漁業	1	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
D 鉱業	11	5	5	0	0	5	5
	100.00	45.45	100.00	0.00	0.00	45.45	100.00
E 建設業	2,096	1,109	1,101	166	166	943	935
	100.00	52.91	99.28	7.92	100.00	44.99	99.15
F 製造業	2,533	1,026	1,000	63	60	963	940
	100.00	40.51	97.47	2.49	95.24	38.02	97.61
G 電気・ガス・熱供給・水道業	4	0	0	0	0	0	0
	100.00	0.00	0.00	0.00	0.00	0.00	0.00
H 情報通信	228	166	164	47	47	119	117
	100.00	72.81	98.80	20.61	100.00	52.19	98.32
I 運輸業	402	176	173	15	15	161	158
	100.00	43.78	98.30	3.73	100.00	40.05	98.14
J 卸売・小売業	243	10	10	1	1	9	9
	100.00	4.12	100.00	0.41	100.00	3.70	100.00
K 金融・保険業	89	52	52	18	18	34	34
	100.00	58.43	100.00	20.22	100.00	38.20	100.00
L 不動産業	1,360	836	833	264	264	572	569
	100.00	61.47	99.64	19.41	100.00	42.06	99.48
M 飲食店，宿泊業	2,524	1,866	1,852	502	501	1,364	1,351
	100.00	73.93	99.25	19.89	99.80	54.04	99.05
N 医療，福祉	161	124	119	42	42	82	77
	100.00	77.02	95.97	26.09	100.00	50.93	93.90
O 教育，学習支援業	306	192	185	59	59	133	126
	100.00	62.75	96.35	19.28	100.00	43.46	94.74
P 複合サービス業	79	54	52	10	10	44	42
	100.00	68.35	96.30	12.66	100.00	55.70	95.45
Q サービス業（他に分類されないもの）	2,490	1,488	1,464	287	286	1,201	1,178
	100.00	59.76	98.39	11.53	99.65	48.23	98.08

別表Ｉ－１－１の脚注参照

別表Ｉ－４　　自動格付と地方格付との一致数（率）及びその正解数（率）
（特定産業（別表Ｉ－７)を除外した結果）

（推薦基準値指定なし）

格付対象事業所数				自動格付数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
								自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
A県	事業所産業	合計	8,387	5,821	4,720	4,650	4,620	1,610	1,609	3,040	3,011
			100.00	69.41	81.09	55.44	99.35	19.20	99.94	36.25	99.05
		新設	5,186	3,857	3,337	3,279	3,259	1,310	1,310	1,969	1,949
			100.00	74.37	86.52	63.23	99.39	25.26	100.00	37.97	98.98
		変動	3,201	1,964	1,383	1,371	1,361	300	299	1,071	1,062
			100.00	61.36	70.42	42.83	99.27	9.37	99.67	33.46	99.16
	企業産業	合計	1,014	823	611	593	591	281	281	312	310
			100.00	81.16	74.24	58.48	99.66	27.71	100.00	30.77	99.36
		新設	180	148	107	97	95	56	56	41	39
			100.00	82.22	72.30	53.89	97.94	31.11	100.00	22.78	95.12
		変動	834	675	504	496	496	225	225	271	271
			100.00	80.94	74.67	59.47	100.00	26.98	100.00	32.49	100.00
B県	事業所産業	合計	42,697	29,841	23,524	22,703	22,552	8,435	8,425	14,268	14,127
			100.00	69.89	78.83	53.17	99.33	19.76	99.88	33.42	99.01
		新設	30,149	22,019	18,370	17,689	17,581	7,260	7,250	10,429	10,331
			100.00	73.03	83.43	58.67	99.39	24.08	99.86	34.59	99.06
		変動	12,548	7,822	5,154	5,014	4,971	1,175	1,175	3,839	3,796
			100.00	62.34	65.89	39.96	99.14	9.36	100.00	30.59	98.88
	企業産業	合計	4,217	3,346	2,355	2,283	2,271	1,135	1,134	1,148	1,137
			100.00	79.35	70.38	54.14	99.47	26.91	99.91	27.22	99.04
		新設	1,027	805	555	516	509	244	243	272	266
			100.00	78.38	68.94	50.24	98.64	23.76	99.59	26.48	97.79
		変動	3,190	2,541	1,800	1,767	1,762	891	891	876	871
			100.00	79.66	70.84	55.39	99.72	27.93	100.00	27.46	99.43

＊ 事業所産業用ルールは、商業統計調査対象の卸売・小売業を含まないルール

（注）１　「格付対象事業所数」は､新設・変動事業所（ただし､事業所産業の場合は商業統計調査対象の卸売・小売業を除く）の総数。

２　「自動格付数」は､格付対象事業所のうち､特定の産業（別表Ｉ－７）及び自動格付されなかったものを除いた数。自動格付率は，格付対象事業所数（特定産業を含む）に対する自動格付数の比率。

３　「自動格付と地方格付との一致数」は、自動格付されたもののうち、自動格付と地方格付が一致した数。ここでいう一致率は、格付対象事業所数（特定産業を含む）に対する一致数の比率

４　「人手審査格付との一致数」は､「自動格付数」又は「自動格付と地方格付との一致数」のうちの「人手審査格付との一致数」であり、いわゆる「正解数」のこと。
また、ここでいう「人手審査格付との一致率」は、いわゆる「正解率」のこと。

５　率の単位は%

別表Ｉ－５　　自動格付と地方格付との一致数（率）及びその正解数（率）
（キーワード（別表Ｉ－８)で除外した結果）

（推薦基準値指定なし）

格付対象事業所数				自動格付数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	全文一致方式		単語分割方式	
								自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
A県	事業所産業	合計	8,387	6,514	5,280	5,195	5,169	1,794	1,793	3,401	3,376
			100.00	77.67	81.06	61.94	99.50	21.39	99.94	40.55	99.26
		新設	5,186	4,359	3,772	3,707	3,688	1,458	1,458	2,249	2,230
			100.00	84.05	86.53	71.48	99.49	28.11	100.00	43.37	99.16
		変動	3,201	2,155	1,508	1,488	1,481	336	335	1,152	1,146
			100.00	67.32	69.98	46.49	99.53	10.50	99.70	35.99	99.48
	企業産業	合計	1,014	872	643	623	622	303	303	320	319
			100.00	86.00	73.74	61.44	99.84	29.88	100.00	31.56	99.69
		新設	180	158	114	102	101	61	61	41	40
			100.00	87.78	72.15	56.67	99.02	33.89	100.00	22.78	97.56
		変動	834	714	529	521	521	242	242	279	279
			100.00	85.61	74.09	62.47	100.00	29.02	100.00	33.45	100.00
B県	事業所産業	合計	42,697	34,372	26,982	26,037	25,881	9,641	9,632	16,396	16,249
			100.00	80.50	78.50	60.98	99.40	22.58	99.91	38.40	99.10
		新設	30,149	25,515	21,166	20,380	20,264	8,262	8,254	12,118	12,010
			100.00	84.63	82.96	67.60	99.43	27.40	99.90	40.19	99.11
		変動	12,548	8,857	5,816	5,657	5,617	1,379	1,378	4,278	4,239
			100.00	70.58	65.67	45.08	99.29	10.99	99.93	34.09	99.09
	企業産業	合計	4,217	3,526	2,458	2,383	2,370	1,201	1,200	1,182	1,170
			100.00	83.61	69.71	56.51	99.45	28.48	99.92	28.03	98.98
		新設	1,027	858	586	546	538	264	263	282	275
			100.00	83.54	68.30	53.16	98.53	25.71	99.62	27.46	97.52
		変動	3,190	2,668	1,872	1,837	1,832	937	937	900	895
			100.00	83.64	70.16	57.59	99.73	29.37	100.00	28.21	99.44

＊ 事業所産業用ルールは、商業統計調査対象の卸売・小売業を含まないルール

（注）１　「格付対象事業所数」は､新設・変動事業所（ただし､事業所産業の場合は商業統計調査対象の卸売・小売業を除く）の総数。

２　「自動格付数」は､格付対象事業所のうち､特定のキーワードを含むもの（別表Ｉ－８）及び自動格付されなかったものを除いた数。
自動格付率は､格付対象事業所数(特定キーワードを含むものを除外せず)に対する自動格付数の比率。

３　「自動格付と地方格付との一致数」は、自動格付されたもののうち、自動格付と地方格付が一致した数。ここでいう一致率は、格付対象事業所数（特定キーワードを含むものを除外せず）に対する一致数の比率

４　「人手審査格付との一致数」は、「自動格付数」又は「自動格付と地方格付との一致数」のうちの「人手審査格付との一致数」であり、いわゆる「正解数」のこと。
また、ここでいう「人手審査格付との一致率」は、いわゆる「正解率」のこと。

５　率の単位は%

別表Ｉ－６　　自動格付と地方格付との一致数（率）及びその正解数（率）
（合成方式による結果）

（推薦基準値指定なし）

格付対象事業所数				自動格付数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）	キーワード（別表２)で除外した結果			
								自動格付数（率）	人手審査格付との一致数（率）	自動格付と地方格付との一致数（率）	人手審査格付との一致数（率）
A県	事業所産業	合計	8,387	8,122	6,490	6,439	6,332	6,658	5,558	5,474	5,439
			100.00	96.84	79.91	76.77	98.34	79.38	83.48	65.27	99.36
		新設	5,186	5,072	4,315	4,287	4,204	4,436	3,876	3,813	3,788
			100.00	97.80	85.07	82.66	98.06	85.54	87.38	73.52	99.34
		変動	3,201	3,050	2,175	2,152	2,128	2,222	1,682	1,661	1,651
			100.00	95.28	71.31	67.23	98.88	69.42	75.70	51.89	99.40
	企業産業	合計	1,014	974	739	723	719	954	728	711	708
			100.00	96.06	75.87	71.30	99.45	94.08	76.31	70.12	99.58
		新設	180	172	130	118	116	170	130	117	116
			100.00	95.56	75.58	65.56	98.31	94.44	76.47	65.00	99.15
		変動	834	802	609	605	603	784	598	594	592
			100.00	96.16	75.94	72.54	99.67	94.00	76.28	71.22	99.66
B県	事業所産業	合計	42,697	41,074	31,317	30,369	29,955	34,876	27,581	26,698	26,545
			100.00	96.20	76.25	71.13	98.64	81.68	79.08	62.53	99.43
		新設	30,149	29,130	23,666	22,881	22,573	25,863	21,545	20,812	20,699
			100.00	96.62	81.24	75.89	98.65	85.78	83.30	69.03	99.46
		変動	12,548	11,944	7,651	7,488	7,382	9,013	6,036	5,886	5,846
			100.00	95.19	64.06	59.67	98.58	71.83	66.97	46.91	99.32
	企業産業	合計	4,217	3,965	2,822	2,747	2,723	3,837	2,732	2,662	2,642
			100.00	94.02	71.17	65.14	99.13	90.99	71.20	63.13	99.25
		新設	1,027	958	667	632	613	924	649	613	598
			100.00	93.28	69.62	61.54	96.99	89.97	70.24	59.69	97.55
		変動	3,190	3,007	2,155	2,115	2,110	2,913	2,083	2,049	2,044
			100.00	94.26	71.67	66.30	99.76	91.32	71.51	64.23	99.76

別表Ｉ－５の脚注参照

別表Ｉ－７　自動格付対象から除外する産業小分類等

（自動格付と地方格付が一致したもののうち，人手審査で誤りと判断される
可能性の高い産業を含む小分類（製造業は全部）及びその除外する主な理由）

産業大分類	産業小分類	自動格付対象から除外する主な理由
E建設業	０６１ 一般土木建築業	事業の内容が「一般土木工事業」、「土木工事一般」等の場合、人手審査で「０６Ａ 土木工事業（舗装工事業を除く）」になる可能性があるため。
Ｆ製造業	（全小分類）	誤りの偏った小分類はみられず、製造業全般に渡って正解率が低いため、全小分類を除外。
Ｌ不動産業	６９４ 不動産管理業	事業の内容が「ビル設備管理」、「ビルメンテナンス業」等の場合、人手審査で「９０４建物サービス業」になる可能性があるため。
M飲食店, 宿泊業	７０Ａ 一般食堂(別掲を除く)	事業の内容が「集団給食」、「給食事業」等の場合、人手審査で「５７Ａ料理品小売業」になる可能性があるため。
	７１３ 酒場,ビヤホール	事業の内容が「串かつ屋」とある場合、人手審査で「７０Ｂ日本料理店」になる可能性があるため。
Ｎ医療，福祉	７５４ 老人福祉・介護事業	事業の内容が「在宅介護サービス」、「訪問介護」、「介護サービス」等の場合、人手審査で「７５Ｄ訪問介護事業」になる可能性があるため.
Ｏ教育,学習支援業	７７３ 学習塾	事業の内容が「幼児教室」、「英会話教室」、「パソコン教室」等の場合、人手審査で「７７Ｎその他の教養・技能教授業」になる可能性があるため。
Ｑサービス業（他に分類されないもの）	８０５ 土木建築ｻｰﾋﾞｽ業	事業の内容が「登記・測量」、「登記申請､測量」等の場合、人手審査で「８０Ｄ他に分類されない専門サービス業」になる可能性があるため。
	８６１ 自動車整備業	事業の内容等に、「自動車の修理販売」、「自動車整備販売」等「自動車販売」と記入がある場合、人手審査で「５８１自動車小売業」になる可能性があるため。

（注）１　事業所産業の自動格付対象から除外するものは、上記表のすべての産業（小分類）及び「Ｆ製造業」。
　　２　企業産業の自動格付対象から除外するものは、上記表のうち「Ｆ製造業」を除くすべての産業（小分類）。

別 表 Ⅰ－8 人手審査で誤りと判断される根拠となるキーワード等一覧表

（自動格付と地方格付が一致した全事業所産業小分類のうち、正解率が９９.２０％未満のもので、同一の訂正符号（人手審査符号）がある場合を検証）

（3－1）

産業大分類	自動格付産業小分類	人手審査後の産業小分類		誤りと判断される根拠となるキーワード等			当該キーワードが含まれる場合、人手審査を必要とする判断理由
		件数［自動＝地方］	件数［自動≠地方］	名 称	事業の内容	取扱い商品	
E 建設業	061 一般土木建築工事業	06A 土木工事業（舗装工事業等を除く）		「一般土木工事業」等「土木工事」プラス「建築工事業」以外の記入の場合に、誤りと判断されていることが多い。			人手審査は、「土木工事」、「建築工事」の組み合わせ、名称、取扱い商品等から総合的に判断しており、判断根拠となるキーワードを特定することが困難なため、産業小分類単位の人手審査が必要と思われる。なお、自動格付と地方格付が不一致の場合に、人手審査は、「総合建設業」、「一般土木建築工事業」（人手修正追加データで 061 に登録）等の記入内容であっても、地方格付を優先し、「06A 土木工事業（舗装工事業等を除く）」を正しいとしている。
		(7 件)	(146 件)				
	081 電気工事業	082 電気通信・信号装置工事業		通信,信号	通信,信号	通信,信号	他に「電気」という単語があると「081 電気工事業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「082 電気通信・信号装置工事業」になる可能性がかなり高い。
		1 件(3 件)	4 件(8 件)				
F 製造業	（全小分類）			判断根拠となるキーワードを、特定することが困難である。			産業小分類単位にみても、判断根拠となるキーワードを特定することが困難なため、全小分類を自動格付の対象から除外する。（企業産業は、自動格付対象から除外しない。）
K 金融・保険業	671 生命保険業	772 職業・教育支援施設		研修，実践	研修,育成,教育		他に「生命保険業」等の単語があると「671 生命保険業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「772 職業・教育支援施設」になる可能性がかなり高い。なお、「研修所」、「研修センター」は、名称キーワードに登録済みである。
		2 件(2 件)	2 件(2 件)				
L 不動産業	694 不動産管理業	904 建物サービス業		メンテナンス	設備管理,メンテナンス，総合管理	設備管理,メンテナンス，総合管理	他に「マンション」等の単語があると「694 不動産管理業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「904 建物サービス業」になる可能性がかなり高い。なお、自動格付と地方格付が不一致の場合に、人手審査は単に「マンション管理」という記入内容であっても、地方格付を優先し、「904 建物サービス業」を正しいとしている。
		1 件(3 件)	1 件(31 件)				
M 飲食店,宿泊業	703 すし店	57A 料理品小売業			製造販売,製造小売		「すし製造販売」又は「すし製造小売」という記入内容の場合、そのほとんどが「703 すし店」に自動格付されているが、人手審査で名称等から判断し、「57A 料理品小売業」になる可能性が高い。
		4 件(4 件)	3 件(3 件)				
	70A 一般食堂(別掲を除く)	57A 料理品小売業		給食	給食	給食	「集団給食」、「給食業」という記入内容の場合、そのほとんどが「70A 一般食堂(別掲を除く)」に自動格付されているが、人手審査で「57A 料理品小売業」になる可能性がかなり高い。なお、「集団給食」、「給食センター」は、人手修正追加データに登録済みである。
		7 件(7 件)	3 件(3 件)				
	70B 日本料理店	711 料亭		割烹，割ぽう	割烹，割ぽう	割烹，割ぽう	「割烹」、「割ぽう」という記入内容の場合、そのほとんどが「70B 日本料理店」に自動格付されているが、人手審査で、他に「日本料理」という単語がないと「711 料亭」に、また、他に「大衆」、「小料理」等の単語があると「713 酒場，ビヤホール」になる可能性が高い。なお、「割烹」、「割ぽう」は、名称キーワードに登録済みである。
		2 件(2 件)	15 件(22 件)				
		713 酒場，ビヤホール					
		3 件(3 件)	3 件(41 件)				

注 1） 件数は、当該キーワードを含む訂正件数であり、( ) 内の件数は総訂正件数，注 2） 企業産業の結果は、事業所産業の結果に準じる。(製造業を除く)

別 表 Ｉ－８ 人手審査で誤りと判断される根拠となるキーワード等一覧表

（自動格付と地方格付が一致した全事業所産業小分類のうち、正解率が９９.２０％未満のもので、同一の訂正符号（人手審査符号）がある場合を検証）

（３－２）

産業大分類	自動格付産業小分類	人手審査後の産業小分類		誤りと判断される根拠となるキーワード等			当該キーワードが含まれる場合、人手審査を必要とする判断理由
		件数［自動＝地方］	件数［自動≠地方］	名 称	事業の内容	取扱い商品	
M 飲食店，宿泊業	713 酒場，ビヤホール	70A 一般食堂(別掲を除く)		食事処，大衆食堂	食事処，大衆食堂		他に「ビール」、「酒」等の単語があると「713 酒場,ビヤホール」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「70A 一般食堂(別掲を除く)」になる可能性がかなり高い。
		3 件(3 件)	11件(33件)				
		70B 日本料理店		串かつ，串カツ	串かつ，串カツ	串かつ，串カツ	他に「ビール」、「酒」等の単語があると「713 酒場,ビヤホール」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「70B 日本料理店」になる可能性がかなり高い。
		16件(23件)	32件(93件)				
N 医療，福祉	754 老人福祉・介護事業	75D 他に分類されない社会保険・社会福祉		「訪問」、「在宅」、「居宅」というキーワードを含む場合、誤りと判断されていることが多い。			単に「介護サービス」等の記入内容であっても、人手審査で、名称等からの総合的判断により、「75D 他に分類されない社会保険・社会福祉」に訂正されているものがかなりあり、また、「754 老人福祉・介護事業」が正しいとされた中にも、「訪問」、「在宅」、「居宅」というキーワードを含むものが多くあるため、産業小分類単位の人手審査が必要と思われる。
		(83 件)	(62 件)				
O 教育,学習支援業	773 学習塾	77N その他の教養・技能教授業		ジュニア，英会話，幼児，パソコン	ジュニア，英会話，幼児，パソコン	ジュニア，英会話，幼児，パソコン	「英会話教室」、「幼児教室」、「パソコン教室」という記入内容の場合、そのほとんどが「773 学習塾」に自動格付されているが、人手審査で「77N その他の教養・技能教授業」になる可能性がかなり高い。なお、「英会話教室」は、名称キーワードに登録済みである。
		18件(22件)	61件(78件)				
	77G 書道教授業	77N その他の教養・技能教授業		ペン，硬筆	ペン，硬筆	ペン，硬筆	「ペン習字」、「硬筆習字」という記入内容の場合、そのほとんどが「77G 書道教授業」に自動格付されているが、人手審査で「77N その他の教養・技能教授業」になる可能性がかなり高い。
		3 件(3 件)	2 件(4 件)				
	77H 生花・茶道教授業	77N その他の教養・技能教授業		フラワーアレンジメント	フラワーアレンジメント		「フラワーアレンジメント」という記入内容の場合、そのほとんどが「77H 生花・茶道教授業」に自動格付されているが、人手審査で「77N その他の教養・技能教授業」になる可能性がかなり高い。
		2 件(2 件)	2 件(5 件)				
P 複合サービス事業	792 事業協同組合（他に分類されないもの）	911 経済団体		酒販組合，同業組合，電気工事工業組合			「酒販組合」、「同業組合」、「電気工事工業組合」という記入内容の場合、そのほとんどが「792 事業協同組合（他に分類されないもの）」に自動格付されているが、「911 経済団体」が正しい。なお、「酒販組合」は、名称キーワードに登録済みである。
		2 件(2 件)	5 件(8 件)				
Q サービス業（他に分類されないもの）	805 土木建築サービス業	80D 他に分類されない専門サービス業		登記	登記	登記	他に「測量」という単語があると「805 土木建築サービス業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「80D 他に分類されない専門サービス業」になる可能性がかなり高い。また、自動格付と地方格付が不一致の場合に、人手審査は単に「測量」という記入内容であっても、地方格付を優先し、「80D 他に分類されない専門サービス業」を正しいとしている。
		6 件(6 件)	7 件(27 件)				

別 表 Ⅰ－8 人手審査で誤りと判断される根拠となるキーワード等一覧表

（自動格付と地方格付が一致した全事業所産業小分類のうち、正解率が９９.２０％未満のもので、同一の訂正符号（人手審査符号）がある場合を検証）

（3－3）

産業大分類	自動格付産業小分類	人手審査後の産業小分類		誤りと判断される根拠となるキーワード等			当該キーワードが含まれる場合、人手審査を必要とする判断理由
		件数［自動＝地方］	件数［自動≠地方］	名 称	事業の内容	取扱い商品	
Q サービス業（他に分類されないもの）	839 他に分類されない生活関連サービス業	606 写真機・写真材料小売業		カメラ店	販売，小売	販売，小売	他に「プリント」、「現像」等の単語があると「839 他に分類されない生活関連サービス業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「606 写真機・写真材料小売業」になる可能性が高い。
		3件(3件)	1件(1件)				
		60F 他に分類されない小売業			販売，小売	販売，小売	他に「ペット」等の単語があると「839 他に分類されない生活関連サービス業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「60F 他に分類されない小売業」になる可能性が高い。
		2件(2件)	0件(1件)				
	861 自動車整備業	581 自動車小売業			販売，小売	販売，小売	他に「修理」という単語があると「861 自動車整備業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「581 自動車小売業」になる可能性がかなり高い。
		31件(32件)	15件(16件)				
	871 機械修理業（電気機械器具を除く）	531 一般機械器具卸売業			販売	販売	他に「修理」という単語があると「871 機械修理業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「531 一般機械器具卸売業」になる可能性が高い。
		2件(2件)	0件(1件)				
	891 広告代理業	899 その他の広告業			折込,新聞,チラシ	折込,新聞,チラシ	他に「広告」という単語があると「891 広告代理業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「899 その他の広告業」になる可能性がかなり高い。
		6件(6件)	8件(11件)				
	90A 労働者派遣業	906 警備業		警備	警備	警備	他に「派遣」等の単語があると「90A 労働者派遣業」に自動格付されやすいが、当該キーワードを含む場合は、人手審査で「906 警備業」になる可能性がかなり高い。
		2件(2件)	1件(1件)				

Ⅱ．　サポートベクターマシンによる産業分類自動格付の研究

岡本　政人*

要　　旨

本稿は，統計センターが平成８年事業所・企業統計調査から部分的に導入している現行産業分類自動格付システムの分類性能を上回る自動格付システムの開発可能性を探るため，平成 13 年事業所・企業統計調査データを用い，第Ⅰ章に述べた実験と同じセッティングでサポートベクターマシン（SVM）による自動格付の実験を行った結果をまとめたものである。

SVM を統計分類自動格付に適用した事例は少ないが，統計分類よりも通常長い文書を対象としているテキスト自動分類の研究では，従来手法よりも大幅に分類性能が向上することが多く，既に一般的となっている手法である。事業所・企業統計調査データを用いた今回の実験では，単純に SVM を適用した場合，現行自動格付システムの使用方法を工夫した合成方式を上回る結果は得られなかったが，現行自動格付システムの格付結果を学習データ・格付対象データに加える add-code 方式（高橋，高村，奥村[2004]）や記入パターン別に区分して適用することなどにより，合成方式を上回る分類性能が得られ，統計分類自動格付の問題に対しても SVM が有効な手法になり得ることが示された。特に，企業産業分類に関しては比較的大きな効果が得られた。しかし，事業所産業分類では期待した程の効果が得られなかったため，今後は，事業所産業分類に対して有効な SVM の適用方法や新たな自動格付法を研究する必要がある。

*　統計センター研究センター
E-mail: research@nstac.go.jp

Ⅱ． サポートベクターマシンによる産業分類自動格付の研究

岡本 政人

はじめに

テキスト自動分類の研究にサポートベクターマシン（SVM: Support Vector Machine）はよく用いられており，ナイーブ・ベイズ法など従来手法に比べて高い分類性能が得られている（平，向内&春野[1998]など）。産業分類や職業分類など統計調査の自由形式の回答を該当する分類区分に自動格付する統計分類の自動格付の問題に対しても，最近になって SVM の適用研究が国内で報告されている（高橋，高村&奥村[2004], ）。外国統計機関においても SVM 適用は考えられているようであるが，オランダ中央統計局の Michiels, & Hacking [2004]は，従来の自動格付法と方法が大きく異なる機械学習法を格付担当職員が受け入れるか懸念があるとして慎重な態度をとっており，具体的な研究結果を報告するに至っていない。

本資料では，平成 16 年事業所・企業統計調査の産業分類格付審査業務への自動格付法の適用可能性の検討に併せて，SVM の適用可能性について研究した結果を説明する。研究結果全体をみると事業所・企業統計調査の産業分類の自動格付の問題に対しても SVM を適用することで分類性能の向上が期待できる。しかし，単純に SVM を適用すると現行自動格付システムの利用方法を工夫した合成方式の分類性能と同程度のため，SVM 適用の効果を得るには従来方式との併用（add-code 方式）及び調査回答の構造に着目して区分ごとに SVM を適用するなどの“工夫”が必要である。今回の研究で採用した“工夫”により，企業産業分類についてはある程度の分類性能の向上が得られたが，事業所産業分類については効果が小さく，今後さらに研究が必要である。

なお，単純な比較はできないが，高橋，高村&奥村[2004]が，日本版総合社会調査 JGSS の職業分類の自動格付に SVM を適用した結果をみると，単純に SVM を適用した場合でも従来のルールベース方式を上回る結果を得ており，統計分類あるいは自動格付の対象となる調査事項などによって適切な SVM の適用方法が異なる可能性がある。

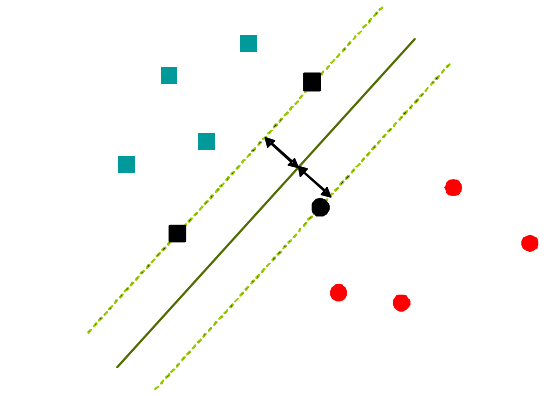
１ サポートベクターマシンの概要

Vapnik [1995]が提案した SVM は，２つのグループを図Ⅱ－１に示すようにマージン最大の（超）平面あるいは（超）曲面で分ける方法である。

マージン最大の（超）平面あるいは（超）曲面に最も近いデータをサポート・ベクターと呼ぶ。

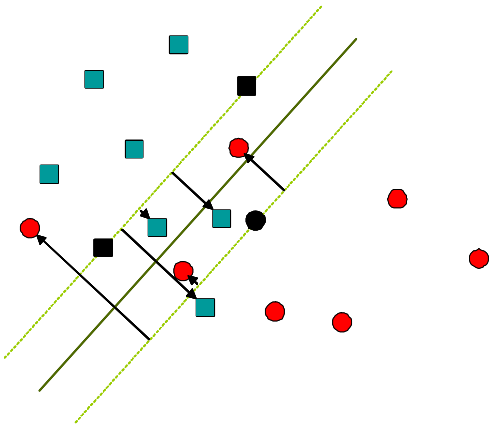
実際の分類問題に適用するには，（超）平面あるいは（超）曲面で完全に区分することは困難であるため，ある程度の分類誤りを許容し，正しく区分できないデータに対してペナルティを課した上でマージン最大の（超）平面あるいは（超）曲面を求めるソフトマージンを用いる。

図Ⅱ－１　２つのグループをマージン最大で区分する直線



■（及び■），●（及び●）は，それぞれ異なる区分に該当するデータ
■，●は，サポート･ベクター

図Ⅱ－２　ソフトマージンを導入したマージン最大化



実際には，区分に用いる（超）平面あるいは（超）曲面の種類（カーネル）及びソフトマージンのペナルティの強さ C_+, C_- を指定し，以下の算式により（超）平面あるいは（超）曲面に対応するパラメータ $\{\alpha_i\}$ を求める。

$$\arg\max_{\{\alpha_i\}}\left[\sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j)\right] \quad \left(\begin{matrix} 0 \le \alpha_i \le C_+ & \text{if } y_i = 1 \\ 0 \le \alpha_i \le C_- & \text{if } y_i = -1 \end{matrix},\ \sum_i \alpha_i y_i = 0\right)$$

ここで，

$\mathbf{x}_i$: 学習用データ（格付ルール生成用データ）

$$y_i = \begin{cases} 1 & \text{if } \mathbf{x}_i\text{が正例（当該グループに属する）} \\ -1 & \text{if } \mathbf{x}_i\text{が負例（当該グループに属さない）} \end{cases}$$

$$K(\mathbf{x}_i,\mathbf{x}_j) = \begin{cases} \mathbf{x}_i'\mathbf{x}_j & \text{：線形カーネル（（超）平面）の場合} \\ (1+\mathbf{x}_i'\mathbf{x}_j)^d & \text{：}d\text{次多項式カーネルの場合} \end{cases}$$

C_+, C_- : それぞれy_iが正，負の場合のソフトマージンのペナルティの強さ

上式で，（超）曲面に対応するカーネルとして多項式カーネルを掲げているが，ガウシアン・カーネルなど他のカーネルも可能である。以下では，２グループのうち $y_i = 1$ としているグループに着目して説明する。当該グループに属する正例を当該グループに属さない

負例として区分してしまう場合のペナルティと，逆に負例を正例に区分してしまう場合のペナルティの強さの比C_+ / C_-をコストファクターと呼んでいる。通常，コストファクターは 1 に設定するが，今回の研究で用いる one-vs-rest 法（後述）のように正例と負例の数が大きく異なる場合 1 と異なる値に設定すると分類性能が向上する可能性がある。負例が正例に区分されてしまう場合のペナルティの強さは$C_- = n / \sum_{i=1}^{n} \mathbf{x}_i' \mathbf{x}_i$とした。

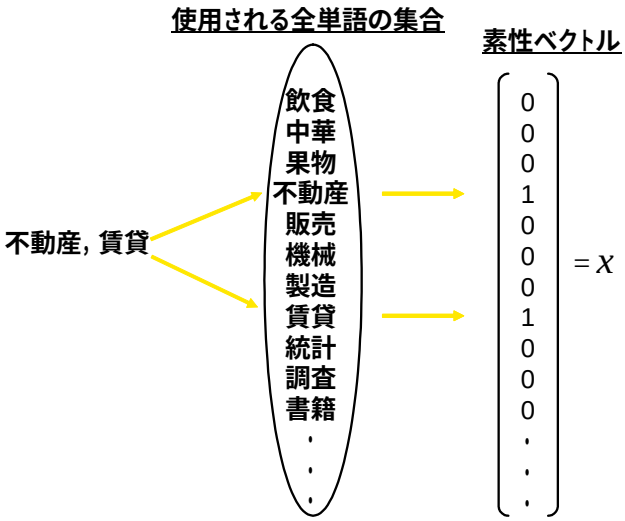
上式によって求めたパラメータから自動格付を行うには，以下の判別関数を用いる。判別関数の値は（超）平面あるいは（超）曲面からの符号付きの距離を示しており，基本的にはこの値が正の場合当該グループに属すると判定する。

$$\sum_i \alpha_i y_i K(x_i, x_j) + b \quad (b : \textbf{バイアス項})$$

グループ（分類区分）が 3 つ以上の場合については，one-vs-rest 法あるいは one-vs-one 法によって複数の 2 値分類の問題に置き換えて SVM を適用する。one-vs-rest 法の場合，各グループについて当該グループと他のグループ全体の 2 値問題として SVM を適用し，このうち判別関数の値が最も大きかったグループに格付する。one-vs-one 法の場合，2 つのグループに限定して SVM を適用し，これを全ての組合せで行い多数決原理で格付する。自動格付では，データ量が多くなるため，通常 one-vs-rest 法が用いられる。今回の研究でも one-vs-rest 法を用いた。

テキストデータに SVM を適用するには，テキストデータを数値データに変換する必要がある。通常，学習用データを形態素解析にかけて分割して得られる単語（不要語は削除）それぞれを座標軸とする非常に次元の高い空間を考え，各テキストデータを形態素解析にかけて分割して得られる各単語の数（bag of words），あるいは，これに適当なウェイトを乗じたものをこの空間上における当該テキストデータの座標とする（図Ⅱ－3 参照）。言語構造をある程度反映させるため，テキスト上で連続している N 個の単語の組（単語 N-gram）を単位にする場合などもある。

図Ⅱ－3　SVM のテキスト自動分類問題への適用方法



2 SVM の適用範囲

一般的なテキスト自動分類と統計分類の自動格付の違いの一つは，処理するテキストの長さが後者のほうが短いことである。1 単語のみの場合も多く，同一テキスト（調査回答）が複数回出現することが少なくない。このため，統計分類の自動格付では，全文一致方式を重視することが多い。ここで全文一致方式とは，学習データにおいて同一テキストが複数回出現し,（統計的に見て）いずれも同じ分類Aに格付されているとみなされる場合に「当該テキスト → 分類A」という格付ルールを設定する方法である。一般的に全文一致方式による格付は十分正確であることから，SVM の適用範囲には含めないこととし，現行自動格付システムで単語分割方式を適用している範囲を SVM の適用範囲とした[2]。

ただし，事業所産業分類の場合，基本的には 2 調査項目「事業の種類」及び「取扱商品」から 1 つのテキストを合成し自動格付を行っており，全文一致方式が適用できる割合は小さいが，第Ⅰ章で述べたように,「事業の種類」の記入が十分詳細であれば,「事業の種類」のみで（自動格付を行い）全文一致方式を適用して十分正確な格付結果が得られることから,「事業の種類」に関して全文一致方式が適用できないテキストを SVM の適用範囲とした。

全文一致方式を適用できないテキストは 2 つに大別した。一つは，学習データにおいて同一テキストの出現頻度が少ないもの，他方は，出現頻度は多いものの同一テキストに対して異なる分類区分に格付されているものである。後者は，例えば「飲食店」のように産業小分類の格付を行うには回答内容が不十分であるが，格付される可能性のある分類区分が限定されるとみられる。今回の研究では，全文一致方式が適用できなかった調査回答をその出現頻度が 5 以上か 5 未満かで分けて別々に SVM を適用している。

今回の研究で用いた自動格付テスト用データ 2 県の新設及び変動事業所[3]のうち, 事業所産業分類では 57%（頻度 5 以上 20%，頻度 5 未満 37%），企業産業分類では 68%（それぞれ 18%，50%）が SVM の適用範囲に当たる。

3 「事業の種類」が全文一致方式ルール非該当・出現頻度 5 以上の場合

1) 企業産業分類

「企業全体の事業の種類」の回答が，学習データにおいて出現頻度 5 以上であるが全文一致方式が適用できない場合, その回答が十分詳細でないことが多く, 事業所としての「事業の種類」及び「取扱商品」から「企業全体の事業の種類」の分類格付を行う傾向がみられる。そこで，形態素解析[4]によって得られた事業所としての「事業の種類」及び「取扱商品」の単語データ（bag-of-words）と地方格付の分類符号を SVM による自動格付の対象

---

[2] 現行の事業所・企業産業分類自動格付システムの単語分割方式は，ナイーブ・ベイズ方式に類似した確率的推論法であるが，出現頻度が十分多い 2～3 の単語の組合せについても格付ルールを自動生成しており，ある程度単語の共起性が考慮されている。（戸井田&瀬谷[1996]，米澤[2000]参照）

[3] 事業の種類が前回調査と同じであった継続事業所は格付対象に含めていない。使用したデータ等については，第Ⅰ章を参照。

[4] 現行の事業所・企業産業分類自動格付システムが搭載している形態素解析ソフトを用いた。以下同様。

データとした[5]。なお，学習用データのうち，複数回出現しない分類区分が付与されたデータはノイズである可能性もあるため除外した。

表Ⅱ－１に示すように，単純に SVM を適用すると，線形カーネル，コストファクター C=1.0 の場合で一致率[6]54.7%，正解率[7]98.7%，二次多項式カーネル，C=0.5 の場合で一致率 56.1%，正解率 98.3%となり，従来方式（「企業全体の事業の種類」から現行自動格付システムにより格付）の一致率 50.9%を上回るが正解率 99.4%を下回っており，事業所情報利用方式（事業所としての「事業の種類」及び「取扱商品」から現行自動格付システムにより格付）の一致率 57.9%，正解率 99.3%，合成方式（従来方式と事業所情報利用方式の単語分割方式による自動格付結果において第一候補の推薦確率が高いほうを採用）の一致率 62.1%，正解率 99.3%をいずれも下回っている。

そこで，現行自動格付システムの格付結果を SVM の学習用データに加える add-code 方式（高橋，高村&奥村[2004]）と，頻度が一定以上の調査回答（以下，記入パターン）については，記入パターンごとに SVM を適用（以下，記入パターン別 SVM）してみた。なお，頻度一定以上の記入パターンについて記入パターンごとに，学習用データのうち複数回出現しない分類区分が付与されたデータはノイズである可能性もあるため除外した。この除外処理の結果，地方格付の分類区分が１つだけになった場合は，SVM を適用せず，この分類区分を付与することとした。また，除外処理の結果，該当事業所がゼロになった場合は，自動格付不可とした。

最も結果が良かったのは合成方式の格付結果を add-code し，記入パターンごとに適用する頻度の下限を 20，２次多項式カーネル，C=1.3 を用いた場合で，一致率 67.6%，正解率 99.7%となり，従来方式，事業所情報利用方式，合成方式（以下，従来方式等と総称する。）を一致率が 5.5 ポイント以上，正解率も 0.3 ポイント以上上回っている。

[5] 企業数が限られるため，現行自動格付システムによる自動格付についても「企業全体の事業の種類」に加え単独事業所や支所である事業所も含めて事業所としての「事業の種類」を学習用データに用いている。「取扱商品」も含めて自動格付を行う場合は，「企業全体の事業の種類」を学習用データから除外している。

[6] 当該自動格付法による分類区分が地方格付と一致した割合

[7] 当該自動格付法による分類区分が地方格付と一致したもののうち審査格付と一致した割合

表Ⅱ－１　企業産業分類自動格付結果
（「企業全体の事業の種類」が全文一致方式ルール非該当で頻度５以上）
該当企業数：966

		従来方式		事業所情報利用方式		合成方式	
		一致率	正解率	一致率	正解率	一致率	正解率
従来方式等		0.509	0.994	0.579	0.993	0.621	0.993
		add-code無し		add-code事業所情報利用方式		add-code合成方式	
		一致率	正解率	一致率	正解率	一致率	正解率
SVM	線形カーネル, C=1.0	0.547	0.987	0.608	0.990	0.622	0.992
	2次多項式カーネル, C=1.3					0.656	0.989
	2次多項式カーネル, C=0.5	0.561	0.983	0.614	0.985	0.638	0.989
下限30*	線形カーネル, C=1.3					0.653	0.997
	線形カーネル, C=1.0					0.653	0.997
	2次多項式カーネル, C=1.4					0.672	0.997
	2次多項式カーネル, C=1.3					0.675	0.997
	2次多項式カーネル, C=1.2					0.674	0.997
	2次多項式カーネル, C=1.1					0.676	0.995
下限20*	線形カーネル, C=1.3					0.654	0.997
	線形カーネル, C=1.0	0.630	0.990	0.655	0.994	0.655	0.997
	線形カーネル, C=0.9			0.654	0.995	0.660	0.997
	線形カーネル, C=0.8			0.655	0.995	0.657	0.997
	線形カーネル, C=0.5			0.643	0.994	0.647	0.995
	2次多項式カーネル, C=1.4					0.672	0.997
	2次多項式カーネル, C=1.3					0.676	0.997
	2次多項式カーネル, C=1.2					0.674	0.997
	2次多項式カーネル, C=1.1					0.675	0.995
	2次多項式カーネル, C=1.0			0.658	0.997	0.671	0.995
	2次多項式カーネル, C=0.9			0.658	0.995	0.667	0.995
	2次多項式カーネル, C=0.5	0.636	0.990	0.651	0.994	0.664	0.994
下限10*	線形カーネル, C=1.3					0.650	0.995
	線形カーネル, C=1.0	0.642	0.990	0.656	0.995	0.650	0.995
	線形カーネル, C=0.9			0.659	0.995	0.655	0.995
	線形カーネル, C=0.8			0.660	0.995	0.660	0.994
	線形カーネル, C=0.7			0.655	0.994	0.656	0.994
	線形カーネル, C=0.5			0.658	0.992	0.653	0.994
	2次多項式カーネル, C=1.1					0.666	0.994
	2次多項式カーネル, C=1.0			0.656	0.997	0.665	0.994
	2次多項式カーネル, C=0.9			0.659	0.995	0.659	0.994
	2次多項式カーネル, C=0.8			0.659	0.995	0.664	0.994
	2次多項式カーネル, C=0.5	0.646	0.992	0.664	0.994	0.667	0.992

* それぞれ頻度が 10, 20, 30 以上の記入パターンについては個別に SVM を適用した場合

2) 事業所産業分類

「事業の種類」の記入パターンが，学習データにおいて出現頻度５以上であるが全文一致方式が適用できない場合，その回答が十分詳細でないことが多く，「取扱商品」も参照して分類格付を行う傾向がみられる。そこで，形態素解析によって得られた「事業の種類」及び「取扱商品」の単語分割データ（bag-of-words）と地方格付の分類符号を SVM の学

習用データとした。企業産業分類の場合と同様に学習用データのうち，複数回出現しない分類区分が付与されたデータはノイズである可能性もあるため除外した。なお，第Ⅰ章に述べたように製造業，及び特定の産業区分については特定キーワードが「事業の種類」あるいは「取扱商品」の回答に含まれている場合，正解率が低くなる傾向がみられるため，特定キーワード等を含む事業所を除外して集計した結果を主に示す。ただし，一致率計算の母数には特定キーワード等を含む事業所を含めている。

表Ⅱ－２に示すように，単純に SVM を適用すると，線形カーネル，コストファクター C=1.0 の場合で一致率 49.5%，正解率 98.8%，２次多項式カーネル，C=1.0 の場合で一致率 50.3%，正解率 98.8%となり，従来方式（「事業の種類」及び「取扱商品」から現行自動格付システムにより格付）の一致率 52.1%，正解率 99.1%，合成方式（従来方式と取扱商品を除外した単語分割方式の自動格付結果において第一候補の推薦確率が高いほうを採用）の一致率 52.7%，正解率 99.2%をいずれも下回っている。なお，取扱商品除外方式（「事業の種類」のみから現行自動格付システムにより格付）は一致率 41.5%，正解率 99.2%となっている。

特定キーワード等を含む事業所を除外しない場合も同様で，単純に SVM を適用すると，線形カーネル，C=1.0 の場合で一致率 55.6%，正解率 98.0%，２次多項式カーネル，C=1.0 の場合で一致率 56.9%，正解率 97.9%となり，従来方式（一致率 59.3%，正解率 98.1%）及び合成方式（一致率 60.1%，正解率 98.0%）の一致率を下回り，正解率も同程度あるいは若干下回っている。取扱商品除外方式は一致率 47.1%，正解率 98.1%となっている。

そこで，企業産業分類の場合と同様に add-code 方式と記入パターン別 SVM を適用してみた。add-code する分類区分として，従来方式あるいは合成方式による格付結果に加え，変動事業所については前回調査である平成 11 年調査の分類区分（ただし，旧産業分類）も用いた[8]。これは，一般的に転業（事業の種類が変動）後と転業前の事業の種類に関連性があり（村田[1995]参照），転業前の事業の種類を補助情報として付加することで分類性能が向上する可能性があるためである。なお，企業産業分類の場合と同様に，記入パターンごとに学習用データのうち複数回出現しない分類区分が付与されたデータはノイズである可能性もあるため除外した。この除外処理の結果，地方格付の分類区分が１つだけになった場合は SVM を適用せず，この分類区分を付与することとした。また，除外処理の結果，該当事業所がゼロになった場合は自動格付不可とした。

この結果を表Ⅱ－２に示す。最も一致率が高かったのは従来方式の格付結果及び前回調査の分類区分を add-code し，記入パターン別 SVM の頻度下限を 20，線形カーネル，コストファクターC=1.0 とした場合で，一致率 56.7%，正解率 99.0%となり，一致率は従来方式等を上回ったが正解率は若干下回った。正解率が最も高かったのは C=0.9 に変更した場合で，一致率 56.2%，正解率 99.2%となり，従来方式等に比べ一致率は 3.5 ポイント以上上回り，正解率は同水準となる。

前回調査の分類区分の add-code は僅かではあるが効果がみられる一方，企業産業分類

[8] 事業の種類が前回調査と同じ事業所は格付対象に含めていない。

の場合とは異なり，合成方式の格付結果の add-code は従来方式（企業産業分類の場合の事業所情報利用方式に対応）の格付結果の add-code を上回る効果が得られない。

表Ⅱ－2　事業所産業分類自動格付結果

（「事業の種類」が全文一致方式ルール非該当で頻度５以上，特定キーワード等除外）

該当事業所数：10153（特定キーワード等を含む）

	取扱商品除外方式		従来方式		合成方式	
	一致率	正解率	一致率	正解率	一致率	正解率
従来方式等	0.415	0.992	0.521	0.991	0.527	0.992

		add-code無し		Add-code従来方式		add-code合成方式		add-code従来方式&前回		add-code合成方式&前回	
		一致率	正解率	一致率	正解率	一致率	正解率	一致率	正解率	一致率	正解率
SVM	線形カーネル, C=1.0	0.495	0.988	0.545	0.990	0.544	0.990	0.543	0.990	0.542	0.991
	2次多項式カーネル, C=1.0	0.503	0.988					0.555	0.991		
下限100*	線形カーネル, C=1.0	0.526	0.990	0.556	0.991	0.553	0.991	0.557	0.991	0.552	0.991
	2次多項式カーネル, C=1.0							0.560	0.991		
下限50*	線形カーネル, C=1.1							0.560	0.992		
	線形カーネル, C=1.0	0.528	0.991	0.561	0.992	0.559	0.991	0.561	0.992	0.558	0.991
	線形カーネル, C=0.9			0.561	0.992			0.562	0.992		
	線形カーネル, C=0.8							0.559	0.992		
	線形カーネル, C=0.5							0.551	0.991		
	2次多項式カーネル, C=1.1							0.563	0.991		
	2次多項式カーネル, C=1.0			0.560	0.992			0.563	0.991		
	2次多項式カーネル, C=0.9							0.561	0.991		
	2次多項式カーネル, C=0.5							0.555	0.990		
下限30*	線形カーネル, C=1.0	0.532	0.990	0.560	0.992	0.557	0.991	0.563	0.991	0.559	0.990
	2次多項式カーネル, C=1.0							0.561	0.991		
下限20*	線形カーネル, C=1.0			0.564	0.990	0.561	0.990	0.567	0.990	0.562	0.990
	2次多項式カーネル, C=1.0							0.566	0.986		
	2次多項式カーネル, C=0.5							0.558	0.986		
下限10*	線形カーネル, C=1.0			0.564	0.989	0.560	0.989	0.566	0.989		

* それぞれ頻度が 10, 20, 30, 50, 100,以上の記入パターンごとに SVM を適用した場合

４　「事業の種類」が全文一致方式ルール非該当・出現頻度５未満の場合

1)　企業産業分類

３章で述べたように「企業全体の事業の種類」の記入パターンごとに SVM を適用することで分類性能の向上がみられたが，記入パターンの出現頻度が少ない場合，学習用データが過少となるため，記入パターンごとに SVM を適用するのは無理がある。そこで記入パターンに替わる効果的な区分法を探す必要がある。

記入パターンの代替として様々なバリエーションが考えられるが，今回の研究では形態素解析にかけて得られる「事業の種類」の記入パターンの単語データの中で，出現頻度が特定の産業区分に集中する傾向のある単語で代替させる方法を採ることとした[9]。

[9] 企業産業分類の自動格付の場合，学習用データは単独事業所及び支所を含めた事業所の「事業の種類」を用いている。

ア　単語の分布集中度を測る基準

各単語の出現頻度が特定の産業区分に集中する度合いを図る尺度として以下の３種類を採用した。

- 当該単語が出現する産業区分の数（sn）

  学習データにおいて当該単語が出現する産業区分の数を用いる。

- 当該単語の産業区分別出現確率のエントロピー（se）

  学習データにおける当該単語の産業区分別出現割合からエントロピーを求め，分布の集中度として用いる。値が小さくなるほど集中度が高くなる。

$$ce \equiv -\sum_i p_i \log p_i \quad (p_i : \textbf{当該単語の産業区分}i\textbf{の出現割合})$$

ここで， $p_i = 0$**ならば** $p_i \log p_i = 0$**とする。**

  各分類区分の大きさの違いを考慮して, 以下の算式を用いることも考えられる。個々の単語の分布の集中度の基準として上記算式よりも必ずしも適切とは言えないが, 企業数が分類区分によって大きく異なる場合，より適切な基準になる可能性がある。なお, 実際に自動格付を行うときには, 前回調査データで学習することになるため, 分類区分の大きさについても前回調査データに基づくものになる。

$$ce \equiv -\sum_i p_i \log p_i / d_i \quad (d_i : \text{産業区分}i\text{の企業数割合})$$

- （エントロピーに基づく）当該単語が出現する産業区分に限定した分布集中度（se1）

  学習データにおける当該単語の産業区分別出現割合からエントロピーを求め，以下の算式により，当該単語が出現する産業区分に限定した分布集中度を用いる。式中のεは，分母がゼロにならないようにするために人為的に付加する値である。この算式は，アメリカ・センサス局の産業・職業分類自動格付システムにおいて，各単語の重要度に応じたウェイトを付与する算式を参考にしている（Appel & Hellerman [1983]参照）。

$$ce1 \equiv \frac{-\sum_{i=1}^{n} p_i \log p_i}{-\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} + \varepsilon} = \frac{ce}{\log n + \varepsilon} \quad \left(\varepsilon = \frac{n}{n+1} \log \frac{n}{n+1},\ n : \text{当該単語が出現する産業区分の数} \right)$$

イ　データの区分方法

「企業全体の事業の種類」の記入パターンによって区分する場合と異なり，単語によって区分する場合，重複を考慮する必要がある。学習用及び格付テスト用データの区分方法は何通りか考えられる。今回の研究では，以下の３通りの区分方法を採用した。

- 排他的区分（ex）

  学習用データにおいて,「事業の種類」の単語データのうち最も集中度の高い単語によって各事業所を排他的に区分する。学習用データにおける区分内の事業所数が一

定数以上のものについては，区分ごとに SVM を適用し，他の区分については一括して SVM を適用する。

- 学習は非排他的－格付は排他的区分（ov-ex）

学習用データにおいて,「事業の種類」の単語データのうち最も集中度の高い単語の一覧を作成し，この一覧に載っている各単語について「事業の種類」の単語データに当該単語が含まれる全ての事業所を区分する。したがって，重複のある区分方法になる。学習用データにおける当該単語の出現頻度が一定数以上のものについては，区分ごとに SVM を適用し，他の区分については一括して SVM を適用する。自動格付は，各格付用データにおける最も集中度の高い単語に対応する SVM 適用結果を用いる。

- 非排他的区分（ov）

学習用データは ov-ex と同じ方法で区分する。自動格付は，各格付テスト用データの単語データのうち一覧に載っている単語に対応する SVM 適用結果の中で最も判別関数の値が高かったものを用いる。

ウ　使用する調査事項

SVM による自動格付に使用する調査事項は，本来的には「企業全体の事業の種類」であり,「企業全体の事業の種類」の回答が産業小分類格付を行うには十分詳細でない場合に事業所としての「事業の種類」及び「取扱商品」等の他の調査事項を用いるべきであろうが，これに沿った自動格付を行うには，出現頻度が少ない「企業全体の事業の種類」の記入パターンについて小分類レベルの格付を行うのに十分詳細であるか否かを判定し，さらに十分詳細でないものについて曖昧さの状況を調べ（例えば，中分類レベルであれば特定の分類区分に格付可能か，いくつかの中分類区分のどれかに格付可能か，多数の中分類区分が格付候補になるか等々），これに基づいて何らかの類似性尺度により記入パターンを区分する必要がある。今回の研究では，時間的制約のため，一律に「事業の種類」及び「取扱商品」を用い，ア及びイで述べた方法で自動格付することとした。

エ　適用結果

表Ⅱ－3 に格付結果を示す。この研究のために利用したソフトウェアの制約[10]のため，単純に SVM を適用した場合の演算ができなかった。頻度下限 200 とした場合の結果から判断すると，SVM を単純に適用した場合，従来方式及び事業所情報利用方式の一致率 41.7%，47.6%を上回り，合成方式の一致率 52.4%と同程度か若干上回るとみられる。正解率も従来方式の 98.5%と同程度か若干上回り，事業所情報利用方式 98.0%，合成方式 98.3%を上回るとみられる。しかし，合成方式に比べて目立った改善は期待し難い。

そこで，現行自動格付システムの格付結果を SVM の学習用データに加える add-code 方式と，出現頻度が一定以上の（基準値最小）単語ごとに SVM を適用してみた。なお，頻

[10] SVM light を用いたが，SVM light の入力ファイルの作成及び SVM light の出力結果の処理に R を用いており，R の制約のため演算ができなかった。

度一定以上の単語については単語ごとに，学習用データのうち複数回出現しない分類区分が付与されたデータはノイズである可能性もあるため除外した。この除外処理の結果，地方格付の分類区分が１つだけになった場合は，SVM を適用せず，この分類区分を付与することとした。また，除外処理の結果，該当事業所がゼロになった場合は，自動格付不可とした。

今回実験した範囲で最も一致率が高かったのは事業所情報利用方式の格付結果を add-code，集中度の基準を sn（分布産業区分数が最小の単語），区分方法を ov（非排他的区分），単語ごとに適用する頻度の下限を 100，２次多項式カーネル，C=0.3 とした場合で，一致率 62.6%，正解率 98.3%となり，従来方式等の一致率を上回るが，正解率については従来方式及び合成方式を下回っている。従来方式等の正解率と同水準か上回る範囲に限定すると，最も一致率が高かったのは上記のパラメータ等のうち add-code するものを合成方式の格付結果に変更した場合で一致率 61.9%，正解率 98.6%となり，従来方式等の一致率を 9.5 ポイント以上と大幅に上回る。なお，合成方式の格付結果を add-code，集中度の基準を se（産業区分別分布のエントロピー最小の単語），区分方法を ov（非排他的区分），頻度の下限を 200，２次多項式カーネル，C=0.3 とした場合に一致率 62.0%，正解率 98.5%となり，一致率がさらに 0.1 ポイント高く，正解率は従来方式と同水準となるが，正解の企業数が上記と同数である。正解率が最も高かったのは，合成方式の格付結果を add-code，集中度の基準を se1（分布している産業区分に限定した集中度最大の単語），区分方法を ex（排他的区分），単語ごとに適用する頻度の下限を 100，線形カーネル，C=1.0 とした場合で，一致率 55.9%，正解率 98.7%となった。ただし，２次多項式カーネルでコストファクターC を 0.3 未満にすることで一致率及び正解率がより高くなるかもしれない。

表Ⅱ－3　企業産業分類自動格付結果

（「企業全体の事業の種類」が全文一致方式ルール非該当で頻度５未満）

該当企業数：2601

				従来方式		事業所情報利用方式		合成方式	
				一致率	正解率	一致率	正解率	一致率	正解率
従来方式等				0.417	0.985	0.476	0.980	0.524	0.983
				add-code無し		add-code事業所情報利用方式		add-code合成方式	
				一致率	正解率	一致率	正解率	一致率	正解率
下限200*	**2次多項式カーネル**, C=0.3	Ex	Sn	0.586	0.980				
			Se	0.578	0.977				
			Se1	0.540	0.986			0.581	0.986
		Ov	Sn	0.599	0.981			0.619	0.985
			Se	0.599	0.981	0.622	0.982	0.620	0.985
			Se1					0.615	0.984
下限100*	**線形カーネル**, C=1.0	Ex	Sn	0.528	0.979	0.573	0.977	0.575	0.981
			Se			0.569	0.980	0.571	0.984
			Se1	0.516	0.985	0.552	0.985	0.559	0.987
		ov-ex	Sn					0.577	0.983
			Se					0.581	0.982
			Se1					0.569	0.981
		Ov	Sn					0.591	0.982
			Se					0.591	0.982
			Se1					0.592	0.981
	線形カーネル, C=0.3	Ex	Sn					0.562	0.981
			Se					0.558	0.981
			Se1					0.552	0.985
		ov-ex	Sn					0.576	0.981
			Se						
			Se1					0.552	0.983
		Ov	Sn					0.596	0.985
			se						
			se1					0.596	0.984
	2次多項式カーネル, C=1.0	Ex	sn					0.612	0.982
			se						
			se1						
		ov-ex	sn					0.606	0.985
			se						
			se1						
		Ov	sn					0.624	0.982
			se						
			se1						
	2次多項式カーネル, C=0.3	Ex	sn	0.576	0.978	0.606	0.977	0.612	0.979
			se					0.603	0.982
			se1					0.570	0.987
		ov-ex	sn			0.605	0.981	0.606	0.983
			se					0.602	0.982
			se1					0.569	0.984
		Ov	sn			0.626	0.983	0.619	0.986
			se					0.618	0.986
			se1					0.612	0.986
下限50*	**2次多項式カーネル**, C=0.3	ex	sn					0.596	0.983
			se					0.594	0.983
			se1					0.564	0.986
		ov	sn					0.619	0.984
			se					0.619	0.985
			se1					0.616	0.984

* それぞれ頻度が 50, 100, 200,以上の（基準値最少）単語ごとに SVM を適用した場合

2） 事業所産業分類

事業所産業分類についても，記入パターン別 SVM の代替として「事業の種類」の記入パターンの単語データの中で，出現頻度が特定の産業区分に集中する傾向のある単語ごとに SVM を適用する方法を採ることとした。「事業の種類」だけでなく「取扱商品」も含めて特徴的な単語を選択する方法も考えられるが，予備的な実験の結果では良い結果が得られなかった。

各単語の産業区分別分布の集中度を測る尺度及びデータの区分方法については，企業産業分類の場合と同様にそれぞれ 3 種類を採用した。SVM による自動格付に使用する調査事項についても企業産業分類の場合と同様に一律に「事業の種類」及び「取扱商品」を用いた。

表Ⅱ－4 に格付結果を示す。単純に SVM を適用した場合，2 次多項式カーネル，C=1.0 とすると，一致率は 41.5%となり，従来方式（「事業の種類」及び「取扱商品」から現行自動格付システムにより格付）の 38.7%，合成方式（従来方式と取扱商品を除外した単語分割方式の自動格付結果において第一候補の推薦確率が高いほうを採用）の 40.9%を上回るが，正解率は 98.1%となり，従来方式及び合成方式の 98.6%を下回る。そこで，単語によって区分せず一括して SVM を適用し，合成方式の格付結果及び変動事業所について前回調査時点の産業区分を add-code し，線形カーネル，コストファクターC=0.9 とすると一致率 42.5%，正解率 98.6%となり，従来方式及び合成方式の正解率と同水準，一致率は 1.6 ポイント以上上回る。カーネルやパラメータを変更すると，一致率は若干上昇するが，正解率が従来方式等を下回ってしまう。単語によって区分し SVM を適用した場合も一致率の上昇幅は小さく，正解率が低下してしまい，結局，一括 SVM 適用を上回る分類性能は得られなかった。このように企業産業分類とは状況が大きく異なっている。

表Ⅱ－4　事業所産業分類自動格付結果

（「事業の種類」が全文一致方式ルール非該当で頻度５未満，特定キーワード等除外）

該当事業所数：19079（特定キーワード等を含む）

				取扱商品除外方式		従来方式		合成方式					
				一致率	正解率	一致率	正解率	一致率	正解率				
従来方式等				0.376	0.987	0.387	0.986	0.409	0.986				
				add-code無し		add-code従来方式		add-code合成方式		add-code従来方式&前回		add-code合成方式&前回	
				一致率	正解率	一致率	正解率	一致率	正解率	一致率	正解率	一致率	正解率
SVM	線形ｶｰﾈﾙ, C=1.5											0.426	0.985
	線形ｶｰﾈﾙ, C=1.1											0.425	0.985
	線形ｶｰﾈﾙ, C=1.0			0.405	0.982	0.417	0.985	0.423	0.985	0.418	0.985	0.425	0.985
	線形ｶｰﾈﾙ, C=0.9			0.404	0.982	0.416	0.984	0.424	0.985	0.418	0.985	0.425	0.986
	線形ｶｰﾈﾙ, C=0.8											0.426	0.985
	線形ｶｰﾈﾙ, C=0.5											0.423	0.985
	2次多項式ｶｰﾈﾙ, C=1.0			0.417	0.981	0.426	0.984	0.430	0.984	0.430	0.983	0.435	0.983
	2次多項式ｶｰﾈﾙ, C=0.5											0.434	0.983
下限500*	線形ｶｰﾈﾙ, C=1.0	ex	sn									0.425	0.985
	2次多項式ｶｰﾈﾙ, C=1.0	ex	sn									0.435	0.983
下限200*	線形ｶｰﾈﾙ, C=1.0	ex	sn									0.425	0.985
			se	0.407	0.981	0.419	0.985	0.426	0.985			0.427	0.985
			se1									0.427	0.985
		ov-ex	sn									0.427	0.984
			se									0.429	0.984
			se1									0.427	0.982
		ov	sn									0.431	0.984
			se									0.431	0.984
			se1									0.431	0.983
下限100*	線形ｶｰﾈﾙ, C=1.0	ex	sn									0.427	0.985
			se	0.413	0.981	0.423	0.984	0.428	0.985			0.430	0.985
			se1			0.424	0.983	0.429	0.984			0.430	0.984
		ov-ex	sn									0.427	0.984
			se									0.429	0.985
			se1									0.428	0.982
		ov	sn									0.433	0.984
			se									0.433	0.984
			se1									0.433	0.983
	2次多項式ｶｰﾈﾙ, C=1.0	ex	sn									0.435	0.982
下限50*	線形ｶｰﾈﾙ, C=1.0	ex	sn									0.425	0.984
			se	0.407	0.981	0.422	0.983	0.428	0.984			0.430	0.984
			se1									0.428	0.984
		ov-ex	sn									0.423	0.983
			se									0.428	0.984
			se1									0.427	0.982
		ov-ex	sn									0.435	0.984
			se									0.436	0.984
			se1									0.434	0.983

* それぞれ頻度が 50, 100, 200, 300, 500,以上の（基準値最少）単語ごとに SVM を適用した場合

5 SVM 適用効果のまとめ

1) 企業産業分類

第３及び４章で述べた企業産業分類の自動格付への SVM 適用結果と自動格付全体におけるその効果を表Ⅱ－５にまとめた。表Ⅱ－１及び３に掲げた SVM 適用結果のうち，それぞれ正解率が従来方式及び合成方式と同水準か上回るものの中で一致率が最も高いものを選択した。特定キーワード等を除外した場合については，同一パラメータ等の対応する結果を用いた。全文一致方式による格付結果を含めた格付データ全体の自動格付結果は，従来方式の一致率 58.9%，合成方式の一致率 66.3%に対し，SVM では 72.1%となり，従来方式に比べ 13.2 ポイント，合成方式に比べ 5.8 ポイント上回っている。なお，SVM の正解率は全体では 99.3%となり，従来方式の 99.4%を若干下回っているが，これは正解率が相対的に低い「企業全体の事業の種類」の記入パターンが全文一致方式非該当で頻度５未満の場合の一致率が従来方式に比べて 20 ポイント以上と大幅に向上したことが影響している。

2) 事業所産業分類

事業所産業分類については表Ⅱ－６にまとめた。表Ⅱ－２及び４に掲げた SVM 適用結果のうち，それぞれ正解率が従来方式及び合成方式と同水準か上回るものの中で一致率が最も高いものを選択した。製造業及び特定キーワード等を含めた場合については，同一パラメータ等の対応する結果を用いた。「事業の種類」のみの全文一致方式による格付結果を含めた格付対象データ全体の自動格付結果は，従来方式の一致率 61.1%，合成方式の一致率 63.0%に対し，SVM では 64.3%となり，従来方式に比べ 3.2 ポイント，合成方式に比べ 1.3 ポイント上回っている。全体の正解率は 99.4%で，従来方式及び合成方式と同水準である。

6 結論及び今後の課題

テキスト自動分類の研究では一般的となっている SVM 法を，統計分類の自動格付に適用する試みはこれまで殆どみられなかったが，最近になって統計分類の自動格付の問題でも SVM 適用が検討されるようになってきた。

本資料に述べたように事業所・企業統計調査の産業分類についても，SVM を適用することである程度の分類性能の向上が期待できることが分かったが，単純な SVM の適用では現行自動格付システムを活用した合成方式に優ることができず，記入パターン別 SVM や，現行自動格付システムの格付結果も利用した add-code 法を用いるなどの工夫を施す必要があった。特に事業所産業分類では SVM 適用効果が小さく，今後事業所産業分類に対してより有効な自動格付法の研究が必要である。

表Ⅱ－５　企業産業分類自動格付結果（総括）

			該当件数	一致率	正解率
全文*一致方式該当	頻度5以上	従来方式 ①	1556	0.917	0.999
		事業所情報利用方式 ②		0.824	0.999
		両方式一致 ③		0.803	1.000
	頻度5未満	従来方式 ④	108	0.731	1.000
		事業所情報利用方式 ⑤		0.667	0.986
		両方式一致 ⑥		0.574	1.000
全文*一致方式非該当	頻度5以上	従来方式 ⑦	966	0.509	0.994
		事業所情報利用方式 ⑧		0.579	0.993
		合成方式 ⑨		0.621	0.993
		記入パターン別SVM，合成方式結果add-code ⑩		0.676	0.997
	頻度5未満	従来方式 ⑪	2601	0.417	0.985
		事業所情報利用方式 ⑫		0.476	0.980
		合成方式 ⑬		0.524	0.983
		基準値最小単語による区分別SVM，合成方式結果add-code ⑭		0.619	0.986
合計	従来方式（①＋④＋⑦＋⑪）		5231	0.589	0.994
	合成方式（①＋④＋⑨＋⑬）			0.663	0.992
	SVM（①＋④＋⑩＋⑭）			0.721	0.993

（特定キーワード等を除外した場合）

			該当#件数	一致率	正解率
全文*一致方式該当	頻度5以上	従来方式 ①	1556	0.842	0.999
		事業所情報利用方式 ②		0.757	0.999
		両方式一致 ③		0.740	1.000
	頻度5未満	従来方式 ④	108	0.639	1.000
		事業所情報利用方式 ⑤		0.565	0.984
		両方式一致 ⑥		0.500	1.000
全文*一致方式非該当	頻度5以上	従来方式 ⑦	966	0.447	0.995
		事業所情報利用方式 ⑧		0.494	0.994
		合成方式 ⑨		0.524	0.994
		記入パターン別SVM，合成方式結果add-code ⑩		0.568	0.996
	頻度5未満	従来方式 ⑪	2601	0.328	0.988
		事業所情報利用方式 ⑫		0.373	0.981
		合成方式 ⑬		0.410	0.984
		基準値最小単語による区分別SVM，合成方式結果add-code ⑭		0.473	0.989
合計	従来方式（①＋④＋⑦＋⑪）		5231	0.509	0.995
	合成方式（①＋④＋⑨＋⑬）			0.564	0.993
	SVM（①＋④＋⑩＋⑭）			0.604	0.995

* 「企業全体の事業の種類」が全文一致方式に該当・非該当で区分

特定キーワード等に該当する企業を含む

表Ⅱ－6　事業所産業分類自動格付結果（総括）

				該当件数	一致率	正解率
全文*一致方式該当	頻度5以上	取扱商品除外方式	①	20968	0.936	0.996
		従来方式	②		0.912	0.996
		両方式一致	③		0.899	0.996
	頻度5未満	取扱商品除外方式	④	884	0.825	0.989
		従来方式	⑤		0.761	0.990
		両方式一致	⑥		0.714	0.990
全文*一致方式非該当	頻度5以上	取扱商品除外方式	⑦	10153	0.471	0.981
		従来方式	⑧		0.593	0.981
		合成方式	⑨		0.601	0.980
		記入パターン別SVM，従来方式結果&前回結果add-code	⑩		0.633	0.983
	頻度5未満	取扱商品除外方式	⑪	19079	0.495	0.972
		従来方式	⑫		0.506	0.970
		合成方式	⑬		0.542	0.970
		一括SVM，従来方式結果&前回結果add-code	⑭		0.564	0.970
合計	従来方式（②＋⑤＋⑧＋⑫）			51084	0.694	0.986
	取扱商品除外全文一致方式＋合成方式（①＋④＋⑨＋⑬）				0.721	0.986
	取扱商品除外全文一致方式＋SVM（①＋④＋⑩＋⑭）				0.735	0.986

（製造業及び特定キーワード等を除外した場合）

				該当#件数	一致率	正解率
全文*一致方式該当	頻度5以上	取扱商品除外方式	①	20968	0.878	0.998
		従来方式	②		0.856	0.998
		両方式一致	③		0.845	0.999
	頻度5未満	取扱商品除外方式	④	884	0.688	0.998
		従来方式	⑤		0.637	0.998
		両方式一致	⑥		0.596	1.000
全文*一致方式非該当	頻度5以上	取扱商品除外方式	⑦	10153	0.415	0.992
		従来方式	⑧		0.521	0.991
		合成方式	⑨		0.527	0.992
		記入パターン別SVM，従来方式結果&前回結果add-code	⑩		0.562	0.992
	頻度5未満	取扱商品除外方式	⑪	19079	0.376	0.987
		従来方式	⑫		0.387	0.986
		合成方式	⑬		0.409	0.986
		一括SVM，従来方式結果&前回結果add-code	⑭		0.425	0.986
合計	従来方式（②＋⑤＋⑧＋⑫）			51084	0.611	0.994
	取扱商品除外全文一致方式＋合成方式（①＋④＋⑨＋⑬）				0.630	0.994
	取扱商品除外全文一致方式＋SVM（①＋④＋⑩＋⑭）				0.643	0.994

* 「事業の種類」が全文一致方式に該当・非該当で区分

製造業，特定キーワード等に該当する事業所を含む

参　考　文　献

平 博順，向内 隆文，春野 雅彦[1998]．Support Vector Machine によるテキスト分類，情報処理学会研究報告 NL128-24，pp.173-180.

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 NL159-9，pp.53-60.

戸井田 幸記, 瀬谷 惠子[1996]. 産業分類の自動格付技法に関する研究, 統計局研究彙報，第 54 号，pp.87-136.

村田 京子[1995]．平成３年事業所統計調査存続事業所の属性変化に係る分析，統計局研究彙報，第 53 号，pp.37-116.

米澤 哲一[2000]．産業分類自動格付システムの７年間（平成４～１０年度）の研究について，統計局研究彙報，第 59 号，pp.61-97.

Appel , M.V. and Hellerman, E. [1983]. Census Bureau experience with automated industry and occupation coding, *Proceedings of the Survey Research Methods Section*, the American Statistical Association, pp.32-40.

Michiels, J. and Hacking, W. [2004]. Compute assisted coding by interviewers, *Proceedings of European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany, 24-26 May 2004.

Vapnik, V. [1995]. *The nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

III． 国内外における統計分類自動格付法の研究動向

岡本 政人*

要 旨

本稿は，統計センターが平成８年事業所・企業統計調査から部分的に導入している産業分類自動格付システムの改善などに資するため，国内外で採用されている統計分類の自動格付法について文献調査を行った結果をまとめたものである。

各国の統計機関が採用している統計分類自動格付システムの調査研究として，スウェーデン統計局の Lyberg & Dean [1992]が知られている。本稿では，Lyberg & Dean [1992]以降の欧米における自動格付システムの発展状況などをまとめている。収集した文献に示されている自動格付システムの分類性能をみると，自動格付法が目立って進歩しているという印象は受けない。Lyberg & Dean が述べている自動格付システムの状況『これまでのところ単純な完全一致方式を採用したシステムが最も成功しており，より複雑なアルゴリズムを採用したシステムは，一般的にそれ程成功していない。』は，残念ながら，今日に至るまでそれ程変化していないように思われる。

しかし，文献調査から今後の研究の方向を示唆する幾つかの研究事例を見出すことができた。例えば，自動格付法を取り入れたより“インテリジェント”な格付支援システムを CAPI/CATI システムに組み込み，分類格付の経験のない調査員による高精度な格付を目指したオランダ中央統計局の研究，サポートベクターマシンなどテキスト自動分類の研究で一般的となってきた手法を応用した国内の研究，曖昧あるいは格付が難しい回答を分離し，これに対応した格付法を適用するなど，統計分類特有の構造を分析し，回答によって適切なアルゴリズムを探求したスペイン統計院やオランダ中央統計局の研究などが挙げられる。

短期間に大幅な性能向上を実現することは困難であろうが，文献調査から統計分類自動格付法改善の研究に取り組む上で有益な情報が得られた。

* 統計センター研究センター
E-mail: research@nstac.go.jp

Ⅲ. 国内外における統計分類自動格付法の研究動向

岡本 政人

はじめに

自動格付法適用の前提として，自由形式の回答がコンピュータ処理できるように入力されている必要がある。欧米では人口センサス調査票などの OCR 入力が 1990 年代以降普及し，これに伴い自動格付システムを導入する国も 1990 年代以降多くなった。また，CAPI/CATI 調査方式の普及は, 調査時点で格付も行う格付支援システムの可能性を開いた。このように入力の問題がある程度解消されたことにより自動格付法や CAPI/CATI＋格付支援システムを採用する国は増えているが，入力された自由形式の回答を処理する自動格付法が Lyberg & Dean [1992]の調査研究以降進歩しているのか不明であった。そこで，統計センターが 1996 年（平成８年）事業所・企業統計調査から部分的に導入している産業分類自動格付システムの改善などに資するため，欧米の統計機関などが採用している統計分類自動格付法について文献調査を行った[11] 。

なお，本稿では「自動格付」を英語の automated coding に対応する語として用いる。「自動分類」と同義とみなしてよいが，例えば，automated coding of industrial classification を「産業分類の自動分類」などと訳すよりも「産業分類の自動格付」と訳すほうが自然であるなど，統計分類の「自動分類」に関して記述する場合，「自動格付」を用いたほうが自然であることが多い。[12]

１．米国人口センサスの産業・職業分類自動格付システム AIOCS

Appel & Hellerman [1983]によると，米国センサス局における最初の自動格付システムは，1967 年経済センサスのために開発された。初期のシステムは，実際の調査データを学習用データとして用い，キーワードを自動抽出する方式であったが，人口センサスを対象とした実験で格付率が低かったため，結局，学習方式を止め，1990 年センサスの集計では OCR を導入し, Hellerman の考案した産業･職業分類自動格付システム AIOCS（Automated Industry and Occupation Coding System）を採用することになった（Scopp, Haley, & Dalzell [2001]）。2000 年センサスでは外部開発の新しいシステムが採用されたが，Hellerman アルゴリズムは，カナダ統計局の自動格付システム ACTR が採用するなど，自動格付システムの発展過程において，重要な位置を占めているため，まず，1990 年センサスで採用された同システムの概要について(1)で説明した後，(2)で 2000 年センサスに採用した自動格付システムの選考過程と，センサスへの適用状況等について説明する。

---

11 統計センターの産業分類自動格付システムについては，戸井田&瀬谷[1996]，米澤[2000]を参照。このほか，統計センターでは，平成 13 年社会生活基本調査から採用したアフターコーディング方式の生活時間調査の集計のために，生活行動分類自動格付システムを導入している（横内[2003]参照）。

12 産業・職業分類などの自動格付について，英語では”Automated coding”と呼ぶことが一般的なようである。この訳語として「自動コーディング」を用いている例もある。ただし，後述する死因分類の自動格付システムの名称に”Automated Classifier”を用いたものがある。

1) Hellerman システムの概要

Hellerman システムは，自動学習による格付ルールの生成やキーワードの抽出などを行わず，職業・産業分類索引を拡張してキーワード（単語）を作成している。分類区分の決定方法は，基本的には，格付対象データ（調査回答）と類似性の最も高いフレーズをデータベースから選択し，そのフレーズに付与されている分類区分を用いる方法である。類似性尺度は，ウェイト付けしたキーワードの一致度である。特定の分類区分で出現頻度が相対的に高いキーワードほど大きいウェイトを付与し，格付対象データとフレーズに共通するキーワードのウェイトを合計する。以下，主に Appel & Hellerman [1983]の説明に基づいて格付手順の概略を示す。

人口センサスの産業・職業に関する調査事項と記入例（下線を引いた箇所）

産業

- For whom did this person work: (勤め先)
  POST（ポスト社）
- What kind of business or industry was it: (事業の種類)
  NEWSPAPER PUBLISHER（新聞社）
- Is it mainly (MFG, WHSLE, RETAIL, OTHER): (粗い産業分類)
  OTHER（製造業，卸売業，小売業，その他）

職業

- What kind of work was this person doing: (仕事の種類)
  SALES PERSON（販売員）
- What was this person's most important duties: (最も重要な職務)
  SELLS ADVERTISING（広告業）
- Was this person an employee for private company, Government employee, self-employed or working without pay: (従業上の地位)
  PRIVATE（民間企業の被用者）

A) 手順

フェーズ１（データ入力及び事前処理）

ステップ１

関連６調査項目（産業に関する２項目及び（粗い）産業分類に関する１項目と，職業に関する２項目及び従業上の地位に関する１項目）を含む調査データを入力

ステップ２

各単語を標準形式に変換する。例えば，複数形を単数形に変換，同義語・略称の変換，

長さが長い単語の不要な末尾の削除等を行う。

ステップ 3

全ての単語について，データベースに登録されているキーワードとマッチングさせる。マッチングしない単語は，スペル訂正ソフトで処理し，それでもマッチングしないものは除外する。

データベースに登録されているキーワードには，産業・職業ウェイト（ステップ 6 で定義）と，当該キーワードが含まれる全てのフレーズへのポインターが付与されている。”a”，”and”などは除外される。なお，Appel & Hellerman [1983]によるとデータベースに収録されているキーワードは約 7200 語となっている。

ステップ 4

“manufacturing”，”wholesale”のように（粗い）産業分類を示唆する単語がないかチェックする。これが（粗い）産業分類に関する調査項目の回答と一致する場合，ボーナス・スコアを付与する。一致しない場合は，フェーズ 2 において，（粗い）産業分類に関する調査項目の回答と，“ 示唆された ” 産業分類の両方について，データベース内のフレーズの中から候補を検索することになる。

ステップ 5

関連調査項目で回答が重複しているものを除外する。

ステップ 6

マッチングを開始する最初の単語を選択する。最初の単語は，最もウェイトの高いキーワードとする。

各キーワードのウェイトは，当該キーワードを含むフレーズの産業区分別出現割合のエントロピーに基づく次式により算出する。特定の分類区分に出現する傾向のあるキーワードほど，ウェイトが大きくなる。

$$H = \frac{E_u - E_W}{E_W - \varepsilon}$$

ここで，

$$E_W = -\sum_{i=1}^{n} p_i \ln p_i$$

p_i :当該キーワードの分類区分iにおける出現頻度の総出現頻度に占める割合

n :当該キーワードが出現する分類区分数

$$E_u = -\sum_{i=1}^{n} \frac{1}{n} \ln \frac{1}{n}$$

$$\varepsilon = \frac{n}{n+1} \ln \frac{n}{n+1}$$ ：分母がゼロになることを防ぐために設定する人為的な値

フェーズ 2（検索及びスコアの計算）

ステップ 1

（粗い）産業分類が小売業でなく，かつ，従業上の地位が政府の被用者でない場合，勤め先の回答と，回答者の居住地域で従業員規模が一定以上の会社名リストと照合する。完全一致する会社名があり，対応する産業分類区分が一意に決まる場合は，無条件でこれを採用する。同じ地域に複数の事業所をもつ会社の場合でも，産業分類区分が同一であれば，無条件で採用する。産業分類区分が複数の場合，従業員数が最も多い産業分類区分を採用する。Gillman & Appel [1994]によると，1990 年センサスにおいて，会社名リストとの完全一致により産業分類が格付された件数は全体の約 6%と少ない。このため，ファジー検索などの導入が検討されたが，誤った照合が多くなってしまい，良い結果が得られなかった。

ステップ 2（If-then-else ルールによる格付）

“School”, ”Hospital”等出現頻度が高い，あるいは，特別な考慮が必要な単語に対しては，データベース検索をせずに If-then-else ルールにより特定の分類区分に格付できるかどうか判定する。”Teacher”のように，産業，職業の両方を格付できる場合もある。

ステップ 3

最初に選定された単語（キーワード）を含む全てのフレーズをデータベース内で検索する。その際，付随情報により，以下の制限の下で不適当と判定されるフレーズを除外する。

- 従業上の地位に関する制限。政府の被用者あるいは事業主などに限定されるフレーズがある。
- （粗い）産業分類に関する制限。データベース内のフレーズに対応する（粗い）産業分類は，（粗い）産業分類に関する項目の回答あるいは産業に関する項目の回答から “ 示唆される ”（粗い）産業分類と一致しなければならない。
- 産業に関する制限。多くの職業分類区分が，特定の産業分類区分のみに分布する。

ステップ 4

ステップ 3 で除外されなかったフレーズに対して，該当調査項目の回答との類似性スコアを計算する。計算方法は，以下の算式を用いる。さらに，異なる調査項目の回答を結合した擬似回答との類似性スコアを計算する。

$$S = M^3 \frac{\sum_{m=1}^{M} H_m}{A_r A_d} \times 100 + Bonus$$

ここで，

M : 一致した単語（キーワード）数

$\sum_{m=1}^{M} H_m$: 一致した単語のウェイトの総和

A_r : 当該項目の回答の実効単語数（*a*や*the*などの単語及び句読点を除外）

A_d : 比較しているフレーズの実効単語数

完全一致が最良であり，この場合，スコアを 2 倍にする。このほか，フェーズ 1 のステップ 4 に記したようなボーナス，同じ単語が複数含まれるフレーズに対するペナルティを加算する。

このようにして算出した各フレーズの類似性スコアが予め設定した閾値を超える場合，採用候補とし，そうでない場合，棄却する。閾値は，分類区分ごとに目標とする正答率を達成するように設定する（Chen, Creecy & Appel [1993]）。

なお，Gillman & Appel [1994]によると，擬似回答によって分類区分を付与した場合，誤分類になる割合がかなり高くなる。

フェーズ 3（分類区分の選択）

フェーズ 2 のステップ 4 で計算された類似性スコアの降順に，採用候補となったフレーズを並べ，スコアの最も高い採用候補のスコアが閾値を上回り，かつ，次の採用候補に比べ一定割合以上であれば，最終的な勝者とする。そうでない場合は，次のキーワードに対して，フェーズ 2 の処理を繰り返す。

結局，採用するフレーズ（勝者）が決まらなかった場合，人手による分類格付を行う。

B) 具体例

次の回答例で自動格付システムの動作を説明する。産業分類については，If-then-else ルールによる格付が行われ，職業分類については，パターンマッチング＋スコア計算による格付が行われた例となっている。

・ 各単語の標準形式への変換を行い，データベース内のキーワードと照合し，ウェイトを付与する。

人口センサスの産業・職業に関する記入例

産業

- For whom did this person work: (勤め先)
  PRIVATE FAMILY（一般家族）
- What kind of business or industry was it: (事業の種類)
  PRIVATE HOME（一般住宅）
- Is it mainly (MFG, WHSLE, RETAIL, OTHER): (粗い産業分類)
  OTHER（製造業，卸売業，小売業，その他）

職業

- What kind of work was this person doing: (仕事の種類)
  BABY SITTER（ベビーシッター）
- What was this person's most important duties: (最も重要な職務)
  CARE OF CHILDREN（子供の世話）
- Was this person an employee for private company, Government employee, self-employed or working without pay: (従業上の地位)
  PRIVATE（民間企業の被用者）

原記入	標準形式	接尾辞コード	産業ウェイト	職業ウェイト
Private	PRIV	13	1	2
Family	FAMIL	42	3	3
Private	PRIV	13	1	2
Home	HOM	3	1	2
Babysitter	BABYSIT	32	3	15
Care	CAR	3	1	1
Of				
Children	CHILDR	49	2	4

- （粗い）産業分類に関する調査項目の回答は"Other"， 異なる（粗い）産業分類を " 示唆する " 回答は無い。
- 削除すべき重複する回答は無い。
- 産業分類については"Family"，職業分類については"Babysitter"を最初の単語として選択する。これは，それぞれの分類について，最も高いウェイトが付与されているからである。
- 従業上の地位は"Private"であるため, 民間企業の被用者に対して有効なデータベース内のフレーズが検索対象となる。

産業分類区分の付与

・ 次の If-then-else ルールにより, 産業分類区分”761 Private Households”が付与される。
  ① 単語”HOM”は，産業分類に関して高出現頻度の単語である。
  ② 単語”PRIV”は，”HOM”によく付随する次の単語（の標準形式）に該当しない。

Aged	Remodel	Elderl
Convalesc	Old	Improvem
Homel	Tour	Nurs
Med	Build	Retard

職業分類区分の付与

・ 単語（キーワード）”BABYSIT”を含むデータベース内の全てのフレーズを抽出すると，”Babysitter, industry 761” (分類符号 406)と”Babysitter, except industry 761” (分類符号 468”)の２つがヒットするが，２つ目のフレーズは，産業分類 761 の場合対象外である。
・ 残った１つのフレーズに対してスコアを計算すると，一致した単語数，回答の実効単語数，データベース内のフレーズの実効単語数いずれも１（M=A_r=A_d=1），一致した単語のウェイトの総和は 15（$\sum_{i=1}^{M} H_m = 15$）となるため，スコアは次の計算により 1500 となるが，完全一致であるため，２倍の 3000 をスコアとする。

$$\text{スコア} = M^3 \frac{\sum_{i=1}^{M} H_m}{A_r A_d} \times 100 = 1 \bullet \frac{15}{1 \bullet 1} \bullet 100 = 1500$$

・ 3000 は, 予め設定した閾値を上回り, 他の採用候補が無いため, 職業分類区分”406 Child Care Workers, Private Households”が選択される。仮に，スコアが閾値を下回る場合は，ウェイトが次に高い”CHILDR”を含むデータベース内のフレーズが抽出され，スコアの計算を行うことになる。

2) 2000 年センサスで採用された新しいシステム

2000 年人口センサスでは，外部開発された自動格付システムを採用することになった。Gillman [2000]によると, 2000 年センサスでは, 新しい産業・職業分類が用いられるため, システムの変更が必要になったが，時間と資源が不足していたことを理由に挙げている。2000 年センサスで採用するシステムの選考過程で，異なる手法を用いた自動格付システムが同一条件で比較評価されており，参考になると思われるので，以下に説明する。

A) 選考過程

Gillman & Appel [1994]は, 外部開発のシステムの比較評価に入る前に行っていた予備的

実験について報告している。その実験では MBR 法(Memory Based Reasoning，最近隣法と同じ)を用いた Thinking Machines 社のシステム（Creecy et al. [1992]）が Hellerman システムをやや上回る性能を示し，最も評価が高かった。MBR 法は，例えば，前回調査データなど，分類格付済みの参照データの中から格付対象データ（調査回答）と最も類似性の高いデータを選び，その参照データの分類区分を付与する方法である。通常，k-Nearest Neighbor 法（k-NN 法），即ち，類似性の高い参照データを k 個選び，その中で最も多い分類区分を採用する方法を採る。Vasconcelos & Turpeinen [1997]は，この時行われたテストを事例研究として採り上げており，MBR 法の場合，特定の産業には特定の職業が多いといった産業と職業の関係（同時分布）を利用できることをメリットとしている。前述のように旧 AIOCS でも，産業と職業の関係をある程度利用しているが，同時分布の情報を完全には利用していない。少なくとも，この点については MBR 法が優っている可能性がある。しかし，MBR 法は，格付対象データごとに膨大な量の参照データとの類似性を計算する必要があるため，処理時間が長くなることが欠点となる。Thinking Machines 社のシステムに対する米国センサス局の評価は高かったようであるが，選考対象には残らなかった。Gillman [2000]は，その理由として，同社が固有の言語でシステムを開発しセンサス局がメインテナンスできなかったこと，テスト結果の検証ができなかったこと，MBR 法がセンサス局の採用してきた人工知能手法と大きく異なることを挙げている。

結局，約 20 システムの中から有望と思われる 5 システムが採用候補として選定され，最終的にはそのうちの 1 システムが採用された。採用候補となった 5 システムは以下のとおりである。

・ Inference Group（オーストラリア）

Precision Data という名称のシステムで，一括処理による自動格付とコンピュータ支援格付の両方の機能が備わっている。シソーラスをベースとして，自然言語処理の技法を用いている。多数のパラメータがあり，自動格付の結果を採用・不採用の基準の変更が容易である。なお，アイルランドでは，1996 年人口センサスの職業分類の自動格付に Precision Data を導入している（Keogh [1998]）。

・ INSEE（フランス）

フランス統計局が 1993 年から開発している SICORE は，bigram（連続する 2 文字）に基づいてツリー形式で自動格付を行うユニークな方法を採っている（３項参照）。

・ National Institute for Occupational Safety and Health (NIOSH)（米国）

NIOSH は，後述の死因分類自動格付システムを開発した国立健康統計センター（The National Center for Health Statistics: NCHS）の関連機関で，SOIC (The Standard Occupation and Industry Coder)という名称のシステムを HGO 社と協同で開発し，死亡証明から人口動態統計を作成する際に，産業・職業分類の自動格付に使用している。1998 年に正式にリリースされている（Marsh & Layne [2001]）。このシステムは，ル

ールベース，完全一致方式，確率的方式の３つの自動格付方式で構成されている。

・ Statistics Canada（カナダ）

カナダ統計局は，後述する ACTR を開発している。ACTR も(旧)AIOCS と同様に Hellerman アルゴリズムをベースとしているが，職業・産業分類以外にも使えるよう汎用型システムになっている。

・ Wise Enterprises（米国）

Dataware Engineering という名称のシステムで，MBR 法を採用している。このシステムは非常に柔軟性があり，多くの自動格付業務に容易に適用可能である。

以上の５システムの評価テストは，同一データセットを用い，正答率[13]の目標値を産業分類 0.90，職業分類 0.87 に設定して行われた。表Ⅲ－１に結果を示すが，時間的制約のため，米国センサス局が採用している産業・職業分類に合わせてチューニングする時間が十分に与えられていない状態でテストを行っており，Gillman [2000]は，各システムの本来の性能が発揮されていない可能性があることに留意が必要としている。

表Ⅲ－１　自動格付システムの格付率の比較１

	Inference	INSEE	NIOSH	Stat Can	Wise	目標 正答率
産業	0.176	0.266	0.418	0.146	0.380	0.90
職業	0.116	0.277	0.327	0.141	0.328	0.87

評価テストの結果，産業分類で最も格付率[14]の高かった NIOSH，職業分類で NIOSH を僅かに上回って最も格付率が高く，産業分類で NIOSH に次いだ Wise Enterprises 社のシステムに絞り，さらに評価テストが行われた。評価テストの最終段階では表Ⅲ－２の結果のとおり，産業，職業分類とも NIOSH が Wise を上回り，1990 年センサスにおける(旧)AIOCS の格付率も上回ったため，NIOSH のシステムを採用することが決まった。ただし，Gillman[2000]は，元来，産業・職業分類の格付業務のために開発されたシステムと汎用的システムを比較しており，Wise Enterprises 社の場合チューニングする時間が十分でなかった可能性もあると述べている。

表Ⅲ－２　自動格付システムの格付率の比較２

	NIOSH	Wise	目標正答率
産業	0.620	0.487	0.90
職業	0.615	0.502	0.87

[13] 自動格付システムが分類区分を付与し，その格付の推定信頼度が予め定めた基準値に等しいか上回ったレコードのうち，付与された分類区分が適切であると判定される割合

[14] 自動格付システムが分類区分を付与した格付対象レコードのうち，その格付の推定信頼度が予め定めた基準値に等しいか上回ったレコード数の総格付対象レコード数に占める割合。

B) 2000 年センサスへの適用状況

新 AIOCS システムを 2000 年センサスに適用した暫定的な結果を表Ⅲ－３に示す。格付率が産業分類で 58.6%，職業分類で 56.0%となり，特に職業分類が 1990 年センサスから改善し，正答率も向上した（Kirk et al. [2001]）。

表Ⅲ－３　AIOCS の格付率と正答率

Production	2000		1990	
	IND	OCC	IND	OCC
Total records/people through autocoder (in millions)	22.5	22.5	N/A	N/A
Gross production rate (assigned a code)	86.4%	80.8%	94.0%	88.0%
Net production rates (gross x acceptance)	58.6%	56.0%	58.0%	37.0%
From validation sample (percent)	94.0%	92.3%	90.0%	87.0%

Gross production rate：自動格付システムが分類区分を付与した割合
Net production rate：格付率，脚注４参照。
Accuracy（”From validation sample”の行）：正答率，脚注３参照
なお，2000 年は暫定的な結果である。

正答率と格付率は閾値によって異なる。閾値を高くすると正答率は上昇するものの格付率は低下する。データの質を考えると当然正答率が高いほうがよいが，格付率が低下すると人手による格付の業務量が多くなり，効率が低下する。米国人口センサスの集計では，分類区分ごとの正答率がなるべく人手格付の場合に近くなるよう分類区分ごとにきめ細かく閾値を設定している（Scopp, Haley & Dalzell [2001]）。

NIOSH の Marsh & Layne [2001]によると，NIOSH の自動格付システムは産業・職業とも人手格付との一致率（＝格付率×正答率）が 87%と高い。単純に比較すると 2000 年センサスへの適用結果と大きく異なるようにみえる。これは，産業・職業分類を死亡証明から格付する場合と，人口センサスの調査票から格付する場合で，調査方式や産業・職業分類の格付の基となる調査事項の違いなどが影響しているかもしれない。

なお, 残念ながら NIOSH のシステムの自動格付法について今のところ詳細が発表されていない。

２．カナダの統計分類自動格付システム ACTR

Tourigny & Moloney [1997]，Wenzowski [1998]によると，カナダ統計局の汎用型自動格付システム ACTR（Automated Coding by Text Retrieval）は 1986 年に開発され，その後改良が加えられている。ACTR の採用している自動格付法は，カナダ統計局における行政記録の照合アルゴリズムを開発した経験と，米国人口センサスの産業・職業分類自動格付システム AIOCS が採用している Hellerman アルゴリズムに基づいている。1991 年の人口センサスでは，人種，母語など比較的簡単な調査項目の自動格付及び対話型格付支援システムとして用いられた。

AIOCS と違う点は，ACTR では，調査回答を標準形式に変換した後，まず参照ファイルに完全一致するものがあるか検索し，完全一致できなかったものについて，AIOCS と同様にキーワードの類似度により格付を行っていることである。カナダ統計局では，それぞれ直接照合（Direct Matching），間接照合（Indirect Matching）と呼んでいる。さらに Gillman & Appel [1994], Wenzowski [1998]の説明から，これ以外の主な相違点が分かる。

- 人口センサスを対象としたバッチ処理だけでなく，CAPI/CATI 方式の調査で調査員が回答を PC に入力する時に自動格付を行う機能を有している。
- 産業・職業分類以外にも利用できるよう汎用型システムとして開発されたため，標準では，産業分類，職業分類を別々に自動格付することになる。AIOCS と同様の処理を行うには，ユーザーがカスタマイズする必要がある。
- データベースを動的に更新できるようになっており，自動格付業務の途中でデータベースに追加できるようになっている。
- 入力データを標準化するなどの事前処理（フェーズ１）の方法は，調査項目に応じて変更できるように設計されている。
- キーワードに付与するウェイトや，回答とデータベース内のフレーズとの類似性を示すスコアの算式は，Tourigny & Moloney [1997]が紹介している ACTR Ver1.06 と Wenzowski [1998]が紹介している Ver3 で異なっている。以前は AIOCS とほぼ同じ算式であったが，その後，以下の式とおり，エントロピーに基づかない算式に変更された。なお，調査項目によって算式を変更することはできない。

$$\text{キーワードのウェイト} = 1 - \frac{\log(\text{回答に含まれるキーワード数})}{\log(\text{総キーワード数})}$$

$$a = \frac{2 \times \text{共通するキーワード数}}{\text{当該レコードのキーワード数} + \text{回答に含まれるキーワード数}}$$

$$b = \frac{\text{共通するキーワードのウェイトの合計}}{\text{回答に含まれるキーワードのウェイトの合計}}$$

$$\text{回答と各レコード間のスコア} = 10 \times \left(\frac{a + 2b}{3} \right)$$

なお，当初は，スペルチェックの機能がなかったが（Gillman & Appel [1994]），現在はスペル訂正機能を有している（Gillman [2000]）。

表Ⅲ－４は，カナダ 1991 年センサスで ACTR を採用することを目指して，同システムの実験を行った結果である（Gillman & Appel [1994]）。”Place-of-Birth”（出身地)や”Ethnic Origin”（人種)のように，回答の範囲が限られる調査項目の格付率は高くなっている。Gillman [2000]によると，ACTR は，１～２単語の非常に短い回答に適した設計になっており，これより長い回答も多い産業・職業のような調査項目の分類性能は必ずしも高くないようである。

Tourigny & Moloney [1997]によると，2001 年人口センサスでは ACTR を産業・職業の自動格付にも適用することを目指しており，ACTR の改良を行っている可能性はあるが，今のところ 2001 年人口センサスへの適用結果について報告を入手していない。

表Ⅲ－４　ACTR の実験結果

VARIABLE	MATCH %	ERROR %
Mother Tongue	92.1	3.4
Place of Birth	91.6	2.0
Ethnic Origin	93.8	1.3
Major Field of Study	78.0	4.4
Industry - Company Name	31.5	8.2
Industry - Kind of Business	38.0	25.5
Industry - Linked Files	22.3	2.4
Occupation - Line 1	42.7	31.1
Occupation - Line 2	19.2	37.0

Matching rate = 格付率(Net production rate)
Error = 1 – 正答率(Accuracy)

なお, イタリア統計院は, 2001 年人口センサスの産業, 職業, 教育程度の自動格付に ACTR を採用している（Macchia & Mastroluca [2004]）。

３．フランスの N-gram による統計分類自動格付システム SICORE

Schuhl [1996]，Rivièra [1997]によると，フランス統計院 INSEE は，QUID という自動格付システムを開発し，1983 年から多くの格付業務に適用してきた。その後，QUID を改良した SICORE (Système Informatique de Codage des Réponses aux Enquêtes)を 1993 年から開発している。SICORE は自動学習によって格付ルールを生成，bigram をベースにしているなど，米国の旧 AIOCS やカナダの ACTR が採用している Hellerman アルゴリズムと大きな違いがある。また，同システムは，どの分類格付にも適用できるよう汎用的シ

ステムになっている。

SICORE の自動格付の方法を以下に示す。

学習フェーズ

・ 正規化ステップ

学習用データから，”of”，”and”など不要な単語を削除，同義語を置き換えるなどの処理をした後，パラメータ指定により，各単語を 1 文字(monogram)，2 文字(bigram)，3 文字(trigram)あるいは 4 文字(quardrigram)ずつ分解する（以下，それぞれを「文字列」と記す)。これらを総称して N-gram と呼ぶ。実際には bigram を用いることが多い。以下に bigram を用いた例を示す。

例　職業分類の学習用データ

TAXI	DRIVER	
DOCTOR		
DOCTOR	OF	MEDECIN
FACTORY	WORKER	
EMPLOYEE		
SURGEON		
SURGEON	AND	DENTIST

正規化済み学習用データ

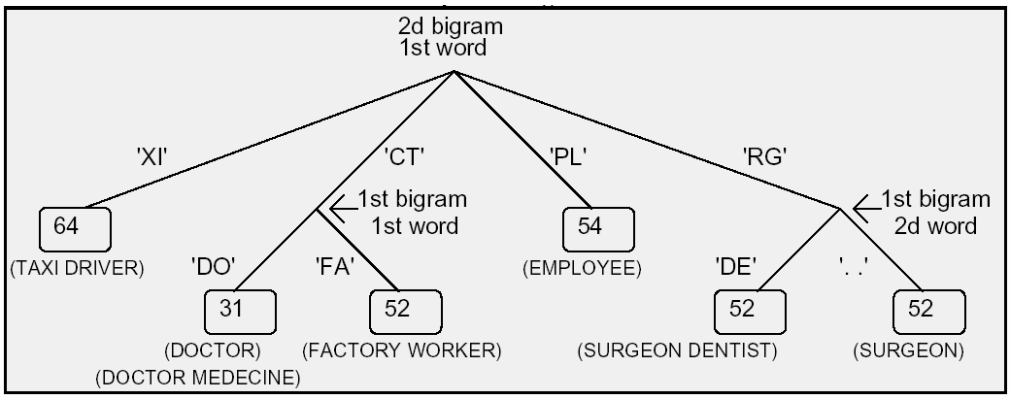
TA	XI				
DO	CT	OR			
DO	CT	OR			
FA	CT	OR	Y		
EM	PL	OY	EE		
SU	RG	EO	N		
SU	RG	EO	N		

DR	IV	ER			
ME	DE	CI	NE		
WO	RK	ER			
DE	NT	IS	T		

・ コーディングツリー生成ステップ

正規化した学習用データ全体から，最大の情報量（シャノン情報量）を与える「文字列」の「位置」を求める。そして，この「位置」の「文字列」によって「分岐」させ，各「分岐」ごとに，次に大きい情報量を与える「位置」をもとめ，さらに「分岐」させる。これを各「分岐」が一つの分類区分に対応するようになるまで繰り返す。こうして生成されたツリーをコーディングツリーと呼んでいる。なお，「位置」を指定してツリーの生成を行うよう指定することも可能である。

コーディングツリーの例



格付フェーズ

- 正規化ステップ
  格付対象データに対して，学習フェーズの正規化ステップと同じ処理を行う。
- パターン認識ステップ
  コーディングツリーを用いて，正規化された格付対象データの格付を行う。結果は，次の３種類に分けられる。
  - 格付できなかった場合。処理はここで終了。
  - 完全に格付できた場合。処理はここで終了。
  - 部分的に格付できた場合。回答（格付対象データ）が曖昧過ぎるため，完全な格付けができない。この場合，論理的ルールが適用される。
- 論理的ルール適用ステップ
  付加的情報を利用して，決定表形式で作成されている論理的ルールによる格付を行う。

自動格付の結果

表Ⅲ－５に調査項目別に SICORE による自動格付結果を掲げる。表Ⅲ－４に示したカナダの ACTR の場合と同様に調査項目によって格付率及び正答率は大きく異なっている。また，同じ統計分類であっても，職業４桁分類についての労働力調査と人口センサスの結果から分かるように，調査によっても格付結果が大きく異なっている。労働力調査は面接調査方式，人口センサスは留置方式であることから，調査方式の違いが格付率に影響しているのかもしれない。

表Ⅲ－５　調査項目別自動格付の結果

Variable	Survey	Efficiency	Accuracy
Occupation, 4 digits (a)	Labor Force Survey	80%	> 90%
	1990 Census	66%	> 90%
Occupation, 2 digits (a)	Survey on living conditions	76%	95%
Occupation, 2 digits (b)	Administrative source[3]	82%	> 95%
Place of vacation	Survey on living conditions	93%	99%
Financial product	Survey on households investments	61%	Good
Human activity (during a day)	Time used Survey	70%	90%
Name and address of establisment	1990 Census	49%	> 90%
Town	1990 Census	94 to 99%	99%

N-gram による自動格付法は，スペルが類似した回答は，該当する分類区分も同じになる傾向があることを前提にしているが，この前提が適切でない場合もある（６の(2)参照）。アメリカ・センサス局の AIOCS システムや単語を単位とした他の自動格付システムと違い，多少のスペルミスや未知語があっても一応格付が可能なことや，ある程度言語の構造を反映できる点が長所と言える。しかし，完全一致による格付に比べると精度は落ちる可能性があるため，完全一致方式と併用することも考えられる。

なお，N-gram の用い方は，SICORE の方式以外にも考えられる。オーストリア中央統計局が 2001 年人口センサスへの適用を目標として開発した自動格付システムは，過去の調査データから得られた分類格付済み参照ファイルの中から最も類似度の高いデータに付与されている分類区分を用いる最近隣法を採用しており，類似度として一致する N-gram の数を用いている。（Haslinger [1997]）

オーストリアのシステムでは，例えば，”VIENNA”は，bigram の場合５つのオーバーラップする bigram {”VI”, “IE”, “EN”, “NN”, “NA”}に分割する。trigram の場合４つのオーバーラップする trigram {”VIE”, “IEN”, “ENN”, “NNA”}に分割する。そして，以下の式により類似度を求める。

$$S(p,q)=1000\frac{|t(p)\cap t(q)|}{\sqrt{|t(p)||t(q)|}},\quad t(p),\ t(q):\text{データ}p,q\text{の}\mathrm{N-gram}\text{の集合}$$

４．オランダにおける機械学習法及びエキスパート・システムの適用研究

オランダ中央統計局は，CAPI/CATI 調査方式を先進的に導入してきたこともあって，後述するコンピュータ支援格付方式の導入に積極的であるが，その一方でナイーブベイズ，

TD-IDF などテキスト自動分類の一般的手法を統計分類自動格付に適用する研究も進めている（Michiels & Hacking [2004]）。

ナイーブベイズ（NB）法は，以下のように各分類区分のサイズ及び各分類区分において当該語が出現する条件付確率から該当する分類区分を推定する方式である。

$$\hat{c} = \arg\max_{c_i} P(c_i)\prod_{k=1}^{n} P(w_k \mid c_i), \quad P(w_k \mid c_i) = (F_{ik} + 0.5)/(N_i + 0.5V_{all})$$

ここで，w_k：格付データに含まれる語

$P(c_i)$＝区分c_iに含まれる学習用データ件数／全学習用データ件数

F_{ik}：区分c_iに語w_kが出現する回数，N_i：区分c_iに出現する語の総数

V_{all}：全学習用データについての語の異なり総数

TF-IDF（term frequency inverse document frequency）法は，以下のように各分類区分において当該語が出現する条件付確率と，当該語を含んでいるデータ件数の全データ件数に占める割合の逆数（特定のデータに集中する傾向を示す一種の尺度）から該当する分類区分を推定する方式である。

$$\hat{c} = \arg\max_{c_i} \sum_{k=1}^{n} P(w_k \mid c_i)\log\frac{D}{D_k}$$

ここで，D：全学習用データ件数，D_k：語w_kを含んでいる学習用データ件数

このほか最近隣（NN）法もテストしている。Michiels & Hacking は，テキスト自動分類の研究で今日では一般的となっているサポートベクターマシン（８の(2)，本資料の第Ⅱ章参照）にも言及しているが，研究対象に含めているとは述べていない。

表Ⅲ－６は，オランダの労働力調査における仕事の種類の回答について連続する３文字（trigram）を語として用い，３方式で職業分類の自動格付を行った結果を示している。ナイーブベイズ法に比べ，TF-IDF 法及び最近隣法の格付率は低くなっている。なお，格付誤りの割合は C5 という決定木方式の格付結果を基準にしているようなので参考にならないかもしれない。

表Ⅲ－６　各種機械学習法による職業分類自動格付のパフォーマンス

Coding technique	Coding rate	Error rate
	%	%
NB	53 (C3G)	28 (C3G)
TFIDF	49 (C3G)	28 (C3G)
NN	45 (C3G)	27 (C3G)

TD-IDF 法は，同じ語が同一テキストに何回も出現する状況を想定した方法であるが，統計分類のようにテキストが幾つかの単語で構成され，単語一つの場合も多い状況でも適しているかどうか確認が必要であろう。TF-IDF 法を採用した韓国の自動格付システムでは，「製造」や「販売」など多くの回答に出現するが，格付上重要な単語に対してウェイトが過小になることが問題点として指摘されている（Kang [2001]）。

格付が難しい少数のデータについては，エキスパート・システムの適用を研究している。例えば，職業が「教師」や「部門の長」といった回答の場合，可能性のある分類区分が多数存在するが，この場合，各単語の分類区分別出現割合から推定する方法は，信頼できる方法ではなく，他の調査項目の回答も考慮して格付する必要がある。そこで，職業名，仕事の種類及び事業の種類の３項目の回答から格付を行うシンプルな推論エンジン及びルールベースが構築された。

表Ⅲ－７は，「教師」及び「部門の長」という回答について格付を行うエキスパート・システムを労働力調査データに適用した実験結果である。

表Ⅲ－７　エキスパート・システムのパフォーマンス

Coding technique	Coding rate	Error rate
	%	%
CACI		
Teacher	81	-
Head	60	-
Expert system		
Teacher	100	39
Head	65	39
CACI + Expert system		
Teacher	100	-
Head	75	-

このエキスパート・システムの格付精度が良いとは言えないが，これは自由形式項目の回答が格付を行うには十分詳細でないことがよくあることも影響しているようである。回答が十分詳細でない場合などに，他の調査項目の回答を補助情報にして論理的格付ルールにより補完する方法は，他の国からも報告されており（アメリカの旧 AIOCS，フランス SICORE，次項のスペインの自動格付法など），オランダ中央統計局が研究しているこのエキスパート・システムもその一種と見ることができる。

調査項目間の論理的関係を記述したルールベースによる自動格付は今のところ限界があり，全面的に採用するのは現実的ではない。通常の自動格付法では精度の高い格付が難しく，調査項目間の論理的関係に基づく格付が可能な範囲に限定してルールベース方式を併用するのが実際的かもしれない。

５．スペイン人口センサスのプレコード方式及び曖昧な回答に対する自動格付法

2001 年スペイン人口センサスでは，産業・職業の調査にプレコード方式が採用された。産業，職業それぞれについて，分類名及び分類符号のリストを配布し，世帯にリストの中から該当するものを探し，調査票に符号を記入してもらう方式である。しかし，小分類レベルで調査するため，リストに全分類区分を網羅できず，リストに載っていない区分につ

いては，自由形式の欄に記入してもらった。また，リストから選んだ分類区分が適切であるか自信がない場合も，同じ欄に記入してもらった。（Jimenez et al. [2003]）

2001 年人口センサスの職業，産業に関する質問

産業	**あなたが働いている事業所または仕事の場所の主な活動は何ですか？** 活動の種類の表（赤字のタイトルがある白い紙）から該当するものを探し，対応する番号を記入； もし，該当する活動の種類が見付けられない，あるいは，疑問がある場合，以下に記入；
職業	**あなたの職業は何ですか？** 注意：あなたの学位（学士，博士・・）や職業上の地位（役人，雇主・・・），労働区分（熟練労働者，見習い・・・）ではなく，仕事の種類を記入してください。 職業の種類の表（黄字のタイトルがある白い紙）から該当するものを探し，対応する番号を記入； もし，該当する活動の種類が見付けられない，あるいは，疑問がある場合，以下に記入；

自由形式欄の回答に対しては，基本的には最近隣法による自動格付を行ったが，例えば，「建設業」のように小分類レベルの格付を行うには十分詳細でない回答も多く，その中には出現頻度の高いものがあるため，出現頻度が 25 を超え，他の調査項目の回答などから格付が可能なものに対しては，通常の自動格付の対象から除外して仮の分類区分（fictitious code）を付与し，論理的ルールによる格付，一部は外部情報に基づく確率的方法で分類区分を付与した。ただし，勤め先の名称などを調査していないため，地域や年齢による傾向などに基づいて実質的に補定を行っている場合も多いようである。仮のコードの数は以下のとおりである。

	産業	職業
仮設コードの数	1,470	988

以下に，プレコード・自由形式欄の記入の有無・自動格付等の可否別件数を示す。なお，（自動）格付可には，プレコードのみ回答があったもの，仮のコードを付与し一種の補定処理を行ったものも含めている。職業の欄に比べて産業の欄の件数が少ないが，これは非回答が多いためである。

		産業			職業		
		格付可	格付否	計	格付可	格付否	計
自由形式欄記入無	プレコード記入有	12,293,295	-	12,293,295	12,409,947	-	12,409,947
自由形式欄記入有	プレコード記入無	1,555,499	500,590	2,056,090	1,501,914	398,153	1,900,067
	プレコード記入有	473,596	130,932	604,527	582,593	137,813	720,406
	計	14,322,390	631,522	14,953,912	14,494,454	535,966	15,030,420

自動格付等の内訳は以下のとおりである。

		産業	職業
自由形式欄の記入による自動格付等	自動格付	1,261,150	1,356,317
	仮のコード	726,110	683,250
	Blank cleaned text?	41,835	44,940
	計	2,029,095	2,084,507

自動格付等の対象データのうち自動格付等ができた割合は，仮のコードを付与したものも含めると下表のとおり 75～80%になるが，通常の自動格付法により格付できた割合は約 50%である。

	産業	職業
自由形式欄記入有	2,660,617	2,620,473
自動格付等可	2,029,095	2,084,507
割合	76.26%	79.55%

プレコードを含めると，全データのうち格付ができた割合は以下のとおり 95%以上になる。プレコード方式は結果に偏りが生じる可能性があり，特にリストに載っていない分類区分が過少になることは避けられないが，省力化という意味では大きな効果が得られたと言える。また，十分詳細でない回答を別途区分（仮のコード）し，区分ごとに他の調査項目などから論理的ルールにより格付を行う方法は参考になるかもしれない。

	産業	職業
回答総数	14,953,912	15,030,420
格付可	14,322,390	14,494,454
割合	95.78%	96.43%

６．死因分類の自動格付・格付支援システム

統計分類の中で死因分類の格付方法は，産業・職業分類などとは異なっており，他の分類にはない複雑さがあると思われるが，これまで開発された自動格付・格付支援システムの格付率及び格付精度をみると，最も成功しているように思われる。

1) 米国の ACME 及び MICAR

Harris [1997]によると，米国の国立健康統計センター（The National Center for Health Statistics: NCHS）は，死亡証明書から死因統計を得るため，死因分類自動格付・格付支援システム ACME（Automated Classification of Medical Entities），MICAR（Mortality Medical Indexing, Classification, and Retrieval）を開発している。

ACME は，まず，死亡証明書に記載されている各死因を人手により国際疾病分類 ICD（International Classification of Disease）符号に格付した後，付与された一連の ICD 符号に対してコンピュータによって WHO ルールを適用し，原死因（直接に死亡を引き起こ

した一連の事象の起因となった疾病もしくは損傷）を選択するものであり，現在では国際的デファクト・スタンダードとなっている。

MICAR は，人手により死因名を ICD 符号に格付していた作業を，ある程度自動化したシステムである。ICD 分類より細かい分類を設定し，各死因について原記入に近い形式で入力させるが，各死因の標準名，その略称又は，直接，数値コードで入力する必要がある。標準名，略称の場合，辞書ファイルを検索して対応する 3～6 桁の数値コードに変換する。この数値コードは，ERN（Entity Reference Number）コードと呼ばれている。さらに，ERN コードで格付した各死因の関連性の論理チェックを行ってから，ICD 符号に自動変換する。

MICAR は，1983 年に開発が始まり，1992 年に実際の利用が認証された。テスト期間終了時には格付率が 85%であったが，今日では 90%になっている。記入単位（各死因）でみると 96%である。MICAR が認証された時点での誤答率（１－正答率）は 0.74%，記入単位でみると 0.42%であった。ちなみに，NCHS の分類担当者（nosologist）の誤答率はそれぞれ 3.5%，2%である。

以上のとおり，MICAR の格付率及び正答率は高い。これは，原記入をそのままを入力するのではなく，人手で標準化等の事前処理を行っていることが一つの要因と考えられる。

なお，Crialesi, R., Frova, L. and Marchetti, S. [1998]は，イタリア統計院が 1995 年データから MICAR-ACME を死因分類格付業務に適用した結果を報告している。同報告によると，MICAR-ACME によって約 80%が自動格付でき，死因統計作成プロセスが改善されたが，人手によって主な死因を選定していた 1994 年以前の統計と断層が生じたため，人手格付と自動格付の一致率（全体では約 90%）を詳細に分析している。

2) スウェーデンの MIKADO

Johansson [1997]によると，スウェーデン統計局では，従来死亡証明書に記載されている各死因を人手により ICD 符号に格付した後，ACME により原死因を選択していたが，各死因を ICD 符号に格付する作業の自動化を進めるため，1993 年にプロトタイプ・システム AKK, 1994 年に MIKADO（MultIpel Kodning Av DödsOrsaker）を開発している。MICAR との相違は，MIKADO の場合，幾つかの標準省略形以外は原記入どおりキー入力し，コンピュータで標準化の処理を行い，該当する ICD 符号に格付できるものは自動格付を行い，格付ができないものについては対話形式でコーダーが格付する方式を採っている点で，MICAR に比べより自動格付システムの性格が強い。

自動格付は，完全一致方式で行っている。これは，死因名に使われる単語は似たようなスペルでありながら意味が大きく異なるものが多く，単語の類似度によって自動格付を行うことが適切でない一方，死因名に使われる用語は比較的少なく完全一致方式が適しているためである。照合率は，記入単位（各死因）でみると標準化の処理を行う前は約 40%であるが，標準化の処理後は 90%を超える。死亡証明書の約 65%は，全記入単位を格付でき，人手によるチェックは不要である。誤答率は，人手格付が 7.2%なのに対し，MIKADO は

3.1%，内訳は，自動格付の誤りが 0.7%，対話形式による格付の誤りが 1.5%，キー入力の誤りが 0.3%となっており，従前の人手格付よりも精度が向上している（正確には，この数字は 1993 年に MIKADO と命名される前のシステムについて検査した結果である。）

７．オランダなどにおけるコンピュータ支援格付方式の研究

今のところ，自動格付システムで 100%格付することができないため，人手による格付を省くことはできない。そこで，人手格付の効率・精度を少しでも向上させるために，自動格付システムとコンピュータ支援型分類格付システムを併用している場合が多い。やや資料が古いが，Lyberg & Dean [1992]によると，オーストラリアやニュージーランド統計局のように，自動格付システムよりもコンピュータ支援型分類格付システムに注力している例もある。オランダ中央統計局の場合も，電話調査あるいは面接調査をしながら調査員がその場でコンピュータに回答を入力する CAPI/CATI 方式を積極的に導入してきた経緯もあり，CAPI/CATI システムに分類格付支援機能を組み込むことにより，分類格付の経験が無い調査員が分類格付を行う可能性が検討されてきた。

CAPI/CATI 方式では，調査時点でデータ入力を行うが，これに自動格付あるいは分類格付支援機能を組み込むことにより，調査時点で格付を行う方法が考えられる。考えらける方式としては，単に印刷されたマニュアルの替りに画面に表示されたマニュアルを参照しながら分類符号を入力させるものから，入力された回答を調査時点で自動格付し，自動格付できなかったものは統計局に回収してから人手格付する方式まで様々考えられる。Gillman & Appel [1994]は, カナダ統計局が開発し税務処理で使用されている階層型メニュー方式の産業分類符号入力システム CASES を観察対象として採り上げ, 訓練された調査員はコンピュータ支援入力方式よりも印刷されたマニュアルを好むとし, 経常人口調査（CPS, 米国の労働力調査）の CAPI/CATI では，格付支援方式よりも自動格付方式が可能か検討しなければならないと記している。このように CAPI/CATI システムに組み込むべき格付方式は各国統計機関の考え方によって異なる可能性もある。

オランダ中央統計局が最近労働力調査で導入した格付支援方式は，調査員が対話形式で入力する CAPI/CATI 方式であることを活かし, 入力された回答が十分詳細かどうかコンピュータが判断し十分でない場合は追加質問を促すなど，自動格付に近い機能を取り入れることなどによって，分類格付の経験のない調査員であっても精度の高い格付を行えることを目指したものとして注目される。

以下，オランダ中央統計局が CAPI/CATI 用ソフトウェア Blaise に組み込んだコンピュータ支援格付方式を用いた研究事例について，自動格付システムと対比する意味で紹介する。

1) Blaise システムの Trigram-Coding

Blaise Version 2.5 で追加された trigram coding (International BLAISE User Group [1993]参照)は，一種の辞書検索機能で，入力データの連続する 3 文字が含まれるフレーズを辞書ファイルから抽出して表示し，その中から調査員が該当する文字列を選択する方法である。例えば，”Stoke-on-Trent”という地名に対応する地域コードを格付するときに，誤って”stroke trent”と入力した場合，辞書に登録されている地名のうち連続する 3 文字が含まれる地名が全て表示される（以下の例を参照)。”stroke trent”と”Stoke-on-Trent”どちらも”oke”が含まれるため，表示される地名には正しい地名”Stoke-on-Trent”が含まれ，これを選択することで，正しい地域コードが入力される。これに対し，アルファベット順の場合，先頭から 3 文字の”Str”まで入力すると，”Stoke-on-Trent”が表示されない。

地名	地域コード
Stoke-on-trent, Stafford	441
Stoke-upon-trent, Stafford	441
Stoke trister, somerset	375
Stradbroke, Suffolk	222
Roke, Oxfordshire	142
Begbroke, Oxfordshire	139
Ladbroke, Warwickshire	452
New bolingbroke, Lincolnshire	514
Old Bolingbroke, Lincolnshire	515
Pembroke, Dyfed	772

2) Trigram-Coding によるコンピュータ支援格付方式の格付率・正答率

Roessingh & Bethlehem [1997]は，オランダの家計調査の品目分類格付に BLAISE の 3 機能を比較テストした結果，アルファベット順方式よりも trigram-coding のほうが良かったが，階層メニュー方式が抜群に良かったと述べている。ただし，この格付業務の経験が十分長く，多くの分類符号を記憶している分類格付担当者を対象としたテストのため，大部分の者がキー入力の少ない階層型メニュー方式をまず選択し，分類が難しく階層メニュー方式で適切な分類区分を見付けられなかったものについてアルファベット順方式か trigram-coding が用いられた。彼らの実験結果から，経験の無い者に格付業務を行わせる場合に trigram-coding が適している可能性があることが示された。

Hardarson [1997]は，アイスランドの個人企業調査と労働力調査の産業分類格付について，調査員に trigram-coding 方式で格付させるテストを行っている。trigram-coding 方式をテストした理由は，労働力調査では，調査データ収集後に分類格付担当者が産業分類格付を行っており，調査員は分類格付の経験が無いためである。

表Ⅲ－８及び９に示すとおり，テストの結果，両調査とも面接調査時に行った分類格付

で 90%超の格付率（hit rate）を達成している。ただし，正答率（Acceptance rate）は分類格付担当者による格付より低くなっており，特に，労働力調査で低くなっている。また，分類格付に係るデータエディティングを含めた総時間の短縮効果は推定で約 8%に止まったため[15]，実際にコンピュータ支援型分類格付方式を採用することについては慎重な結論を導いている。この結果から判断すると，労働投入量の削減という観点では，自動格付システムの効果のほうが大きいかもしれない。なお，表頭の 5digits，4digits，2digits は，それぞれ産業の５桁分類，４桁分類，２桁分類のことである。

表Ⅲ－８　コンピュータ支援型分類格付（Interactive coding）の実験結果

1996 年個人企業調査試験調査

	Interactive coding			Manual coding		
	5 digits	4 digits	2 digits	5 digits	4 digits	2 digits
Acceptable	196	197	205	217	220	227
Unacceptable	10	9	5	5	5	2
Insufficient information	7	7	3	0	0	0
Not coded	18	18	18	9	6	2
Total number of cases	231	231	231	231	231	231
Hit rate[1]	92.2	92.2	92.2	96.1	97.4	99.1
Acceptance rate[2]	95.1	95.6	97.6	97.7	97.8	99.1
Success rate[3]	87.7	88.2	90.0	93.9	95.2	98.3

[1] The number of assigned codes over the total number of cases
[2] The number of accepted codes over the number of assigned codes
[3] Hit rate times acceptance rate

表Ⅲ－９　コンピュータ支援型分類格付の実験結果

1996 年 11 月労働力調査

	ISAT-95 4 digits	ISAT-95 2 digits		
	Total	Total	Inexperienced interviewers	Experienced interviewers
Hit rate	92.5	92.5	93.8	90.5
Acceptance rate	88.7	92.7	91.6	94.3
Success rate	82.0	85.7	85.9	85.3
Number of cases	610	610	357	253

3) オランダ中央統計局の新しいコンピュータ支援型調査員格付方式

オランダの労働力調査では，最近新しいコンピュータ支援格付方式を産業，職業，教育程度の分類格付が導入された。新方式は，単に辞書から該当するものを検索するだけでなく，回答によって検索する辞書あるいは検索方法を切り替えたり，コンピュータが即時に該当する分類区分とその確率を推定し，回答が十分詳細でない場合は追加の質問を行えるようにするなど，自動格付システムに近い機能を搭載している。これにより，これまで統計局において行っていた格付業務を調査員が面接時に行う方式に変更している（Michiels & Hacking [2004]）。

[15] アイスランドは 1992 年から労働力調査などに CAPI 方式を導入しており，調査時点で産業分類格付を行わない場合でも，分類格付に必要なデータ（回答）は PC に入力される。CAPI 方式と紙と鉛筆による調査方式を比較しているのではないことに注意。

職業分類の場合は，以前の調査から得られた「仕事の種類」の回答から単語の分類区分別出現頻度の検索ファイルが用意され，「仕事の種類」の回答を調査員がラップトップコンピュータに入力すると，該当する分類区分を探索する。例えば，“ Carpenter ”（大工）が回答であるとすると，検索ファイルから“ 大工 ”を条件とする各分類区分の条件付き確率を算出，“ Carpenter at shipyard ”（船大工）のように複数の単語から成る回答であると，以下のように単語ごとの条件付確率から各分類区分の確率（ウェイト）を算出し，確率が一定以上のものが複数ある場合は回答者にどれが該当するか選択させる。確率が一定以上の分類区分が無かった場合は，追加情報を得るためさらに質問を行うことができる。

```
Carpenter: P(code1|Carpenter)=0.60, P(code2|Carpenter)=0.20,・・・
At:        P(code14|At)=0.02, P(code2|At)=0.01, P(code11|At)=0.01,・・・
Shipyard:  P(code2|Shipyard)=0.50, P(code4|Shipyard)=0.35,・・・
-------------------------------------------------------------------------
Carpenter+At+Shipyard: Weight(code2)=0.71, Weight(code1)=0.60,・・・
```

産業分類の場合は，２つの検索ファイル：①従業員 100 人以上の会社の名簿と②以前の調査から得られた事業の種類の回答における単語の分類区分別出現頻度が用意されている。まず従業者規模を質問し，回答が 100 人以上であれば，会社名を質問し，①のファイルから類似度の高い会社を検索する。該当する会社が複数で 7 以下の場合，回答者にどれが該当するか選択させる。該当する会社が 7 を超える場合は，事業の種類を質問する。この質問は，従業員が 100 人未満の場合の最初の質問と同じであり，②のファイルに基づいて該当する分類区分を求める。この手順は職業分類と同様である。該当する分類区分が 7 未満か，推定確率が一定以上の分類区分が 6 の場合，回答者にどれが該当するか選択させる。該当する分類区分が無かった場合あるいは多過ぎる場合は，ファイル検索を行わず，階層的な質問形式により，１桁ないし２桁分類のレベルで格付を行う。

教育程度の場合は，正規教育の名前（education name）とそのレベル，分野，機関名といった特性の入った検索ファイルが用意される。教育程度については，以前の調査データを利用していない。通常，教育の名前の回答だけでは該当する候補が多過ぎるため，特性について追加質問を行う。候補が 2～6 に絞られると，回答者にどれが該当するか選択させる。

この方式の格付率及び格付精度は表Ⅲ－１０のとおりである。格付率は７～８割，格付誤りが１割程度となっている。回答者に確認までしている割には格付精度が悪いようにも思えるが，これは採用している分類が産業 415 区分，職業 2,137 区分，教育程度 966 区分と細かいことも影響しているかもしれない。しかし，産業分類については会社名から格付できた場合，事業の種類について質問を省略するため確認情報が限られ，少なくとも形式的には格付精度が良いと推測されることや，実際に会社名から格付できる割合が一定割合

存在すると推測されること（従業員数 100 人以上の会社の従業者は対象人口の 57%を占める）などから，事業の種類から格付できた場合の精度がかなり劣っている可能性があり，今のところ新方式に対して明確な評価をし難い。

表Ⅲ－１０　新しいコンピュータ支援型調査員格付方式のパフォーマンス

Attribute	First phase: September 2003		Second phase: January 2004	
		%		%
Number of respondents	699	100	11501	100
respondents age: 14 years or older	564	81	9128	79
respondents with a job	364	52	5726	50
Number of businesses to be coded	383	100	5978	100
coded with search engine	315	82	4653	78
coded with hierarchical question	49	13	947	16
not coded	19	5	378	6
Number of occupations to be coded	364	100	5726	100
coded with search engine	289	79	4299	75
not coded	75	21	1427	25
Number of (current) educations to be coded	107	100	1650	100
coded with search engine	88	82	1235	75
not coded	19	18	415	25
Number of completed educations to be coded	1199	100	19127	100
coded with search engine	989	83	15569	81
not coded	210	17	3558	19
Error rate economic activity of businesses		<12		7
Error rate occupations		<10		-
Error rate educations		<13		-

調査員が面接時に分類格付を行うことで面接時間がどれだけ長くなったかを表Ⅲ－１１に示す。調査員によるコンピュータ支援格付方式を導入した 2004 年 1 月は，面接時間が 1 件当たり+26 秒長くなったが，その後面接時間は短縮し 3 月には従来との差は+1 秒に止まっている。これは，調査員が操作に慣れてきたことと，職業及び教育程度に係る面接時間が長くなった反面，産業に係る面接時間が短縮したことが影響している。産業の場合，従来は，会社名，所在地及び事業の種類を必ず質問していたのに対し，新方式では会社名のみで格付が可能であると，他の質問はスキップするため，その分面接時間が短縮されている。

表Ⅲ－１１　面接に要した時間

Questionnaire	Interview length (old)	Interview length CACI		
	Labour force survey 2003	Labour force survey January-March 2004		
		January	February	March
Business	72 sec	53 sec	48 sec	49 sec
Occupation	36 sec	49 sec	46 sec	47 sec
Education	177 sec	209 sec	196 sec	190 sec
Total	285 sec	311 sec	290 sec	286 sec

８．国内における研究 － 意味解析及びサポートベクターマシンによる産業・職業分類自動格付の研究

1) 意味解析による自動格付法

高橋[1998，2002]は，「格フレーム」の概念に基づいたルールベースの自動格付システムを開発し，「健康と階層」調査及び「JGSS（Japanese General Social Survey，日本版総合社会調査）」第１回本調査における産業・職業分類格付に適用している。

「格フレーム」は，文の意味を表現する方法として用いられ，動詞を基準として，取り得る格とその値に関する制約を記述したものである。格の種類は，目的に応じて適切なものを用いる。Fillmore は次のような格を考察した（京都大学工学部情報学科計算機科学コースウェア「自然言語処理」（http://www.kuis.kyoto-u.ac.jp/isle/le4-lang/lang.html）参照）。

動作主（A）	与えられた動作を引き起こすもの
経験者格（E）	与えられた心理現象を体験するもの
道具格（I）	与えられた出来事の原因となるもの
対象格（O）	移動や変化する対象物
源泉格（S）	対象物の移動や変化における起点
目標格（G）	対象物の移動や変化における終点
場所格（L）	与えられた出来事が起こる場所
時間格（T）	与えられた出来事が起こる時間

高橋[1998]によると，一般に職業は大まかには動作（述語により表現される）の違いにより分類され，さらに，動作の対象や動作を行う場所などにより細分類される傾向があることから，職業に関する知識の多くは次のような格フレームにより表現できると考え，この形式で格付ルールを定義した。格付データは，意味解析処理により，調査回答を格フレームによる意味表現に正規化し，格付ルールと照合して自動格付を行った。実際に開発されたシステムには，「営業，販売，布団打ち直し」，「住宅の設計・建築」，「米・麦を作る」といった３種類の並列表現の処理機能などもある。産業分類にも同様の方法が適用された。

599	農耕・養蚕作業者の場合
	述語：栽培
	対象格：野菜

522	中学校教員の場合
	述語：教える
	場所格：中学校

「健康と階層」調査及び「JGSS」第１回本調査の産業・職業分類格付への適用結果は，産業分類で正答率 90～95%，再現率（＝格付率×正答率）70～75%，職業分類で正答率 75～85%， 再現率 60～70%となっている。産業分類の自動格付結果が職業分類に比べて良いのは，職業分類が約 200 区分と小分類レベルなのに対し，産業分類は約 20 区分と粗い区分であることなどが影響していると思われる。職業分類の正答率が低いという印象を受けるが，採用している職業分類の難度が本質的に高いのかもしれない。

また，官庁統計では大量データを処理するため，推定精度が一定以上の自動格付結果については人手審査をせずにそのまま集計に使用することを前提として自動格付システムを開発しているのに対し，高橋[2002]によると，このシステムは，従来３回繰り返していた人手格付のうち，１回を自動格付に置き換えることを目的としており，自動格付システムに求められる正答率の許容範囲が広いとみることもできる。

2) サポートベクターマシンによる自動格付法

高橋，高村&奥村[2004]は，職業分類自動格付システムの性能をさらに向上させるには，前述のルールベースによる自動格付法では以下のような限界があると指摘し，サポートベクターマシン（SVM）の適用を研究している。

- 職業分類は，自由形式の「仕事の内容」を中心に「従業先の事業の種類」（自由形式），「従業上の地位」，「役職」，「従業先の事業規模」などの他の調査項目の回答も含めて総合的に判断し該当分類区分を決める必要があるが，これらを全てルールとして表現することは非常に困難である。
- 回答が商品名や生産物のみの場合など，ルールを格フレームの形式で表現できないものもある。（ある調査では約 20%がこれに該当）
- 「仕事の内容」に出現する用語や表現の仕方が時代と共に変化しており，システムの辞書及びシソーラスの更新を継続的に行ったとしても時間的な遅れが伴う。

「JGSS」第３回調査（2002 年）のデータを格付データとして用いた実験結果は，同じデータで SVM を訓練させる 10 分割の交差検定方式（Cross Validation）の場合で再現率は 74.7%，第１～２回調査（2000～2001 年）のデータで訓練させた場合で 71.9%となり，従来のルールベース方式の約 66%を上回った。さらに従来方式によって推定した分類区分を訓練用データ及び格付データに加えた add-code 方式を採ると，SVM の再現率はそれぞれ 76.9%，73.1%と向上した。

この結果は，テキスト自動分類の研究で既に一般的となっているサポートベクターマシンが統計分類の自動格付の問題においても，従来の方式よりも分類性能が優れている可能性があることを示しており，注目される。

おわりに

本稿では，Lyberg and Dean [1992]の調査研究以降の統計分類自動格付法の発展状況を把握するように努めた。米国センサス局が内部開発した非自動学習型の Hellerman システムを維持更新が困難なことを理由に，外部開発の自動学習型システムに切り替えたことに象徴されるように，過去の調査データなどを用いて自動学習させる方式あるいは過去の調査データなどを参照ファイルとして最近隣法を適用する方式へと方法論に変化が見られるものの，収集文献に示されている自動格付システムの分類性能をみると，入力データの標準化を徹底させた死因分類を別にすると，自動格付法が目立って進歩しているという印象を受けない。Lyberg & Dean [1992]が述べている自動格付システムの状況『これまでのところ単純な完全一致方式を採用したシステムが最も成功しており，より複雑なアルゴリズムを採用したシステムは，一般的にそれ程成功していない。』は，残念ながら今日に至るまでそれ程変化していないように思われる。

しかし，自動格付法の研究の進むべき方向が全く見えないという訳ではない。オランダ中央統計局（Michiels & Hacking [2004]）のように，自動格付法を取り入れたより“インテリジェント”な格付支援システムを CAPI/CATI システムに組み込み，分類格付の経験のない調査員による高精度な格付を目指した研究は一つの方向を示している。また，インターネットの普及もあって膨大な文書が電子化されるようになり，テキスト自動分類の研究が盛んになってきたことも明るい材料と言える。実際，同分野で既に一般的となっているサポートベクターマシンを適用することにより，ある程度分類性能が向上する可能性が示されている（高橋，高村&奥村[2004]）。スペイン統計院（Jimenez et al. [2003]）やオランダ中央統計局（Michiels & Hacking [2004]）などのように曖昧あるいは格付が難しい回答を分離し，これに対応した格付法を適用するなど，各統計分類特有の構造を分析し，回答によって適切なアルゴリズムを探求することも一つの方向であろう。

短期間に大幅な性能向上を実現することは困難であろうが，今回の文献調査から統計分類自動格付法改善の研究に取り組む上で有益な情報が得られた。

参　考　文　献

Appel, M.V. and Hellerman, E. [1983]. Census Bureau experience with automated industry and occupation coding, *1983 Proceedings of the American Statistical Association*, Survey Research Methods Section.

Chen, B., Creecy, R.H. and Appel, M.V. [1993]. Error control of automated industry and occupation coding, *Journal of Official Statistics*, Vol. 9, No. 4, pp. 729-745.

Creecy, R.H., Masand, B.M., Smith, S.J., and Waltz, D.L. [1992]. Trading MIPS and memory for knowledge engineering, *Communications of the ACM*, Vol. 35, No. 8, pp. 48-64.

Crialesi, R., Frova, L. and Marchetti, S. [1998]. The impact on mortality related procedures and data of introducing automated cause codingin Italy, *paper presented for NTTS '98*, Eurostat.

Dumičić, S. [1997]. Automated coding in the '91 census in Croatia, *Statistical Data Editing*, Vol.2, United Nations, pp. 209-216.

Evers, T. [2000]. *Progress Report*, 14th International Roundtable on Business Survey Frames, 2000.

Gillman, D. and Appel, M.V. [1994]. Automated coding research at the Census Bureau, *SRD Research Report RR* 94/04, 10/5/94.

Gillman, D. [2000]. Developing an industry and occupation autocoder for the 2000 census, *2000 Proceedings of the American Statistical Association Annual Meeting*, Governemnt Statistics Section.

Hardarson, O.S. [1997]. Interactive coding of economic activity using trigram search in BLAISE III, *Individual Paper*, IBUC 4th Annual International Blaise Users Conference, 1997.

Harris, K.W. [1997]. Evaluation of an automated multiple cause of death coding system, *Working Paper* No. 23, UN/ECE Work Session on Statistical Data Editing 1997.

Haslinger, A. [1997]. Automatic coding and text processing using N-grams, *Statistical Data Editing*, Vol.2, United Nations, pp. 199-209.

International BLAISE User Group [1993]. *Newsletter* No. 2, 1993.

Jimenez, F. H., Serra, F. F., Alvarez, A. A. and Gaviria, A. P. [2003]. Treatment of the economic activity and the occupation in the census of population: Spanish experience, *Working Paper* No. 8, UN/ECE Work Session on Statistical Data Editing, Madrid, 2003.

Kalpić,D. [1994]. Automated coding of census data, *Journal of Official Statistics*, Vol. 10, No. 4, pp. 449-463.

Kang, Y.G. [2001]. AIOC system, *Proceedings of the 53rd ISI Session*, Seoul, Korea

Keogh, G. [1998]. Automatically coding occupation descriptions from the 1996 census of population of Ireland, *paper presented for NTTS '98*, Eurostat.

Kirk, M., Buckles, E., Mims, W., Appel, M.V. and Johnson, P. [2001]. Preliminary results from the census 2000, industry and occupation coding, *2001 Proceedings of the American Statistical Association Annual Meeting*, Governemnt Statistics Section [CD-ROM].

Johansson, L.A. [1997]. Automatic coding of diagnosis expressions, *Statistical Data Editing*, Vol.2, United Nations, pp. 216-221.

Lyberg, L. and Dean, P. [1992]. Automated coding of survey responses: An international review, *R & D report* 1992:2, Green series, Statistics Sweden.

Macchia, S. and Angelis, R.D. [1998]. Applying automated coding to the pilot survey of next population census: Challenge, *paper presented for NTTS '98*, Eurostat.

Macchia, S. and D'Orazio, M. [2001]. A system to monitor the quality of automated coding of textual answers to open questions, *Research in Official Statistics*, Vol.4-2.

Macchia, S. and Mastroluca, S. [2004]. The automatic coding process in the 2001 Italian general population census: efficacy and quality, *conference paper for European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany, 24-26 May 2004 [CD-ROM].

Marsh, S.M. and Layne, L.A. [2001]. Fatal injuries to civilian workers in the United States, 1980-1995, DHHS(NIOSH) Publication No. 2001-129.

Michiels, J. and Hacking, W. [2004]. Compute assisted coding by interviewers, *conference paper for European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany, 24-26 May 2004 [CD-ROM].

Rivièra,P. [1997]. SICORE – general automatic coding system, *Statistical Data Editing*, Vol.2, United Nations, pp. 222-231.

Schuhl, P. [1996]. SICORE, The INSEE automatic coding system, *1996 Annual Research Conference Proceedings*.

Scopp, T., Haley, K., and Dalzell, D. [2001]. A preliminary look at the effects of optical character recognition (OCR) and keying on the quality of industry and occupation coding in census 2000, *2001 Proceedings of the American Statistical Association Annual Meeting*, Governemnt Statistics Section [CD-ROM].

Roessingh, M. and Bethlehem, J. [1997]. Trigram-coding in the family expenditure survey in Statistics Netherlands, *Statistical Data Editing*, Vol.2, United Nations, pp. 180-186.

Tourigny, J.Y. and Moloney, J. [1997]. The 1991 Canadian census of population experience with automated coding, *Statistical Data Editing*, Vol.2, United Nations, pp. 186-198.

Vasconcelos, N. and Turpeinen, M. [1997]. Case studies in memory-based reasoning and learning, *http://web.media.mit.edu/~mtu/mab/nuno/*

Wenzowski, M.J. [1998]. Advances in automated and computer assisted coding software at Statistics Canada, *Individual Paper*, IBUC 5th Annual International Blaise Users Conference, 1998.

高橋 和子[1998]．格フレームによる自由回答のコーディング自動化システム，情報処理学会研究報告，FI51-12，NL127-12，pp. 87-94.

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, 情報処理学会研究報告, NL147-8，pp. 47-53.

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 NL159-9，pp.53-60.

戸井田 幸記，瀬谷 惠子[1996]．産業分類の自動格付技法に関する研究，統計局研究彙報，第 54 号，pp.87-136.

横内 宏至[2003]．形態素解析等の言語処理手法を用いた生活行動分類自動格付システムの開発　－平成 13 年社会生活基本調査－」，統計センター 製表技術参考資料，1.

米澤 哲一 [2000]．産業分類自動格付システムの７年間（平成４～１０年度）の研究について，統計局研究彙報，第 59 巻，pp. 61-97.