機械学習による自動格付システムの開発

下野　寿之†
和田　かず美†
床　裕佳子†

An Autocoding System with a Supervised Multiclass Learning

SHIMONO Toshiyuki
WADA Kazumi
TOKO Yukako

本稿では、短い自由記入の文章を分類する多クラス分類器の開発と、この多クラス分類器を家計調査の収支項目分類に適用する取り組みについて紹介する。この分類器は、教師あり機械学習アルゴリズムによるもので、シンプルで仕組みの理解や実装がしやすく、学習も分類も高速で、正解率が高く、格付けした分類の「信頼度」も併せて算出することができる。この信頼度は、業務の省力化を図る上で非常に重要な役割を果たすことが見込まれる。

キーワード：多クラス分類器、教師あり学習、自動格付、自然言語処理

We have developed an autocoding system for the Family Income and Expenditure Survey in Japan. The proposed system is a supervised multiclass classifier using a machine learning algorithm. The system classifies short text descriptions into multiple classes and have the following advantages. Firstly, as the structure of the classifier is simple, it is easy to develop as a business system. Next, the processing time for labeled training data and making prediction for unlabeled data are practically short compared to a deep learning system provided by a software vendor. In addition, the classifier yields high accuracy results for a large portion of a dataset. Furthermore, it outputs the most promising class label with “trust-value” that indicates the reliability of the outputted label. This trust-value would play an important role in practical use to save labor.

The classifier would be applicable other coding tasks and greatly contribute to improving the efficiency in the field of official statistics.

Key words: Multiclass classifier, Supervised learning, Autocoding, Natural language processing

† 独立行政法人統計センター統計情報・技術部統計技術研究課

1. はじめに

調査統計の調査票に自由記入欄がある場合、その記入内容を統計的に処理するためには、所与の分類体系に基づく分類符号の付与（格付け）を必要とする。従来、専門の職員により行われてきたこの作業は、統計調査の集計事務の中でも非常に多くの人手と時間を要するため、統計センターでは、結果公表までの期間の短縮とデータ処理にかかる費用の削減を目指し、近年、格付処理の自動化に向けて様々な検討が行われている。油井（2017）によれば、自動格付システムは、大きく分けてルールベース方式と機械学習方式があり、職員が人手格付を行う際に得られた知見をルール化し、そのルールに基づき自動格付を行うルールベース方式が主流である。

本稿では、調査ごとに詳細なルールを組み込んだシステムを開発する必要のあるルールベース方式に比べ、システムの構成がシンプルで汎用性の高い機械学習アルゴリズムを用いた多クラス分類器の開発及び、この分類器の家計調査への適用について紹介する。開発した分類器は、ナイーブベイズに近いアルゴリズムを採用しており、その仕組みは本稿で比較を行ったディープラーニングシステムよりもシンプルで理解しやすく、システム開発も容易であり、高速で、正解率が高く、格付けした分類の「信頼度」も併せて算出することができる。この信頼度は、ナイーブベイズによる分類器の事後確率よりも計算が簡単である。

家計調査は、家計簿の自由記入欄に記録された個々の収入・支出の内容を、約 550 項目ある収支項目分類に分類・格付けを行い集計しているが、これまで調査票が冊子形式の家計簿であるため、データの電子化が困難であり、自動格付の導入を妨げていた。このような経緯を踏まえ、公的統計の精度向上に向けた総務省の取り組みとして、家計調査のオンライン化が検討され、平成 30 年 1 月からスマートフォンやタブレットでも記入できるオンライン家計簿が導入されることになり、本研究はこの動きを踏まえた取組みの一環である。

Toko et al. (2017) は、Perl で開発された本稿の提案システムを R で実装し、家計調査のデータを用いて、上述の約 550 項目を統合した 11 項目の分類体系について適用した。CART 及びランダムフォレストとの比較では、提案システムの R 実装版はランダムフォレストと匹敵する精度であることがわかっている。CART は提案手法の R 実装版と同程度に高速であるが精度が悪く、一方でランダムフォレストは計算時間がかかり、家計調査への適用という観点では実用は難しい。また、Toko et al. (2017) は、英語の短文への適用も試み、良い結果が得られている。

第 2 節では、提案システムがアイデアを得たナイーブベイズによる分類器について述べ、その問題点を整理する。第 3 節は、提案する分類器の概要について説明し、ナイーブベイズ分類器との相違点を明らかにする。第 4 節では、実際の家計調査データへの適用により、そのアルゴリズムについて解説する。第 5 節で提案する分類器を用いた格付システムの性能評価を行い、第 6 節でディープラーニングによるシステムとの比較結果について簡単に紹介する。第 7 節に、得られた結論と今後の課題を整理し、第 8 節で公的統計作成における機械学習によるシステムの実用化について考察する。

2. ナイーブベイズによる分類器とその問題点

ここでは、Tsubaki et al. (2017) に基づき、特徴抽出をユニグラムとする場合のナイーブベイズによる分類器について説明し、Toko et al. (2017) が指摘する問題点について整理する。

符号全ての集合を $K = \{1, \dots, M\}$、学習プロセスにおいて抽出された特徴全ての集合を $F = \{F_1, \dots, F_J\}$ 、f_j は$\{0, 1\}$ の値をとるランダム変数で、これによりある特徴の有無を表す。ある符号 k についてある特徴 F_j が存在する確率を $p_{kj} = p(F_j = 1|K = k)$ として、符号 k に分類されるデータがある特徴の組み合わせを持つ場合の条件付確率は、

$$P\left(F_j = f_j;\, j = 1,\dots,J \middle| K = k\right) = \prod_{j=1}^{J} p_{k_j}^{f_j}\left(1 - p_{kj}\right)^{1-f_j} \tag{1}$$

と表現される。このとき、符号 k についての学習用データ内のレコード数を n_k 、同様に符号 k についてある特徴 F_j のレコード数を n_{kj}とすると、p_{kj} の最尤推定量は

$$\hat{p}_{kj} = n_{kj}/n_k \tag{2}$$

により得ることができるが、結果が 0 または 1 になることを避けるために加法スムージングを採用し、任意の定数 α と β を分子分母の両方に加えて次のような形にする。

$$\hat{p}_{kj} = (n_{kj} + \alpha)/(n_k + \beta) \tag{3}$$

事前確率 $P(K = k) = p_k$ とすると、ベイズの定理により事後確率は、

$$P\left(K = k \middle| F_j = f_j;\, j = 1,\dots,J\right) = \frac{P\left(F_j = f_j;\, j = 1,\dots,J \middle| K = k\right) \cdot P(K = k)}{P(F_j = f_j;\, j = 1,\dots,J)}, \tag{4}$$

さらに右辺分母の役割は規格化なので、左辺は右辺の分子と比例する。つまり、

$$\begin{aligned} P\left(K = k \middle| F_j = f_j;\, j = 1,\dots,J\right) &\propto P\left(F_j = f_j;\, j = 1,\dots,J \middle| K = k\right) \cdot P(K = k) \\ &= p_k \prod_{j=1}^{J} p_{k_j}^{f_j}\left(1 - p_{kj}\right)^{1-f_j} \end{aligned} \tag{5}$$

となる。式 (5) から、ナイーブベイズによる分類器は、ある所与の特徴の組み合わせのレコードがあるとき、事後確率が最も高い符号を付与することがわかる。そして、事前分布はサイズ（レコード数）n の学習用データを用いて、$p_k = n_k/n$ とすることが多い。

ところが、Copas (1983) が指摘するとおり、しばしば学習用データは予測したいデータと同じ母集団を代表する標本であるとは限らず、そのような場合に予測に深刻な偏りが起こる。そしてその偏りは、結果から原因を推定する逆推定の問題に関して顕著である。そして、本研究で取り組む家計簿の記入文字列と分類符号の関係性は、逆推定の問題に該当する。つまり、文字列が符号を決めるのではなく、記入対象の商品・サービス等とそれに対応する符号が所与であり、それによって特定の特徴が発生していると考えることができる。

3.　提案する分類器

Tsubaki et al. (2017) は、田口 (1997) の感度を用いて、符号 k を支持する特徴 F_j の感度 β_{kj} を、条件付確率 p_{kj} のロジットとして

$$\beta_{kj} = \log\left(\frac{p_{kj}}{1 - p_{kj}}\right) \tag{6}$$

と定義した。これを用いて、式 (1) の条件付確率の対数をとり、

$$\begin{aligned} \log P\left(F_j = f_j;\, j = 1,\dots,J \middle| K = k\right) &= \log\left\{\prod_{j=1}^{J} p_{k_j}^{f_j}\left(1 - p_{kj}\right)^{1-f_j}\right\} \\ &= \sum_{j=1}^{J} \log\left(1 - p_{kj}\right) + \sum_{j=1}^{J} f_j \log\left(p_{kj}/(1 - p_{kj})\right) \end{aligned}$$

$$= \sum_{j=1}^{J} \log(1 - p_{kj}) + \sum_{j=1}^{J} f_j \beta_{kj} \tag{7}$$

となるので、再び指数化すると、

$$P(F_j = f_j;\ j = 1, \ldots, J \mid K = k) = \exp\left\{\sum_{j=1}^{J} \log(1 - p_{kj}) + \sum_{j=1}^{J} f_j \beta_{kj}\right\}$$
$$= \prod_{j=1}^{J} (1 - p_{kj}) \cdot \exp\left(\sum_{j=1}^{J} f_j \beta_{kj}\right),$$

この $P(F_j = f_j;\ j = 1, \ldots, J \mid K = k)$ から事後確率の分母となる $P(F_j = f_j;\ j = 1, \ldots, J)$ は、

$$P(F_j = f_j;\ j = 1, \ldots, J) = \sum_{l=1}^{M} P(K = l) \cdot P(F_j = f_j;\ j = 1, \ldots, J \mid K = l)$$
$$= \sum_{l=1}^{M} \left\{p_l \prod_{j=1}^{J} (1 - p_{lj}) \cdot \exp\left(\sum_{j=1}^{J} f_j \beta_{lj}\right)\right\},$$

これにより、事後確率は

$$P(K = k \mid F_j = f_j;\ j = 1, \ldots, J) = \frac{\prod_{j=1}^{J} (1 - p_{kj}) \cdot \exp(\sum_{j=1}^{J} f_j \beta_{kj}) \cdot p_k}{\sum_{l=1}^{M} \left\{p_l \prod_{j=1}^{J} (1 - p_{lj}) \cdot \exp\left(\sum_{j=1}^{J} f_j \beta_{lj}\right)\right\}},$$

ここで、さらに事前確率を $p_k \propto \{\prod_{j=1}^{J} (1 - p_{kj})\}^{-1}$ と定義すると、

$$P(K = k \mid F_j = f_j;\ j = 1, \ldots, J) = \frac{\exp(\sum_{j=1}^{J} f_j \beta_{kj})}{\sum_{l=1}^{M} \exp\left(\sum_{j=1}^{J} f_j \beta_{lj}\right)} \tag{8}$$

を得ることができる。

Tsubaki et al. (2017) は、式 (7) の事前確率が無情報の状態で $1/M$となり、$M = J = 2$ のとき、事後確率 (8) が 1 から田口の標準化誤差率を引いたものとなることに着目した。これは、式 (8) に基づく分類器が、標準化誤差率の最大値を最小化していることを示しており、さらには、標準化誤差率を M と J に一般化していると指摘している。

本稿で提案する分類器は、式 (8) をさらに簡素化したもので、Toko et al. (2017) はこれを次のように説明している。まず、

$$\underset{l}{\operatorname{argmax}}\, P(K = l \mid F_j = f_j;\ j = 1, \ldots, J) = \underset{l}{\operatorname{argmax}} \prod_{\{j \mid f_j = 1\}} \frac{p_{lj}}{1 - p_{lj}} \tag{9}$$

という性質に着目する。提案する分類器は、p_{kj} の最尤推定量として、符号のクラス別の確率に着目する式 (3) の代わりに特徴に着目し、

$$\hat{p}_{kj} = (n_{kj} + \beta)/(n_j + \alpha) \tag{10}$$

を採用した。この $\hat{p}_{kj}$ を特徴と符号別の擬似信頼度とする。ここで、$n_j = \sum_{l=1}^{M} n_{lj}$ は、同じ特徴 F_j について符号別の頻度 n_{kj} を合計したものである。ナイーブベイズによる分類器の場合は、式 (3) の分母には、学習用データにおける符号 k の出現頻度を使用していることに留意されたい。

ある特徴 F_j について、符号 k が最有力候補であれば、$\hat{p}_{kj}$ は 1 に近くなるという性質を利用して、任意の入力レコード i から得られた特徴の集合 $J(i)$ についての擬似信頼度 $p_i(\cdot)$ は、

$$p_i(k) = \max_{j \in J(i)} \frac{n_{kj} + \beta}{n_j + \alpha} \tag{11}$$

となる。この $p_i(k)$ に対応する符号 k が、特徴 F_j についての最有力候補である。

次に、特徴毎に得られた候補符号から、次の条件を満たすものを最終的に付与する符号とする。

$$k_i := \arg\max_k p_i(k) \tag{12}$$

この符号とともに出力する信頼度 v_i は、

$$v_i := p_i(k_i) \times \frac{p_i(k_i)^\gamma}{\sum_k p_i(k)^\gamma} \tag{13}$$

と定義する。ここで、γ は調整用パラメータである。

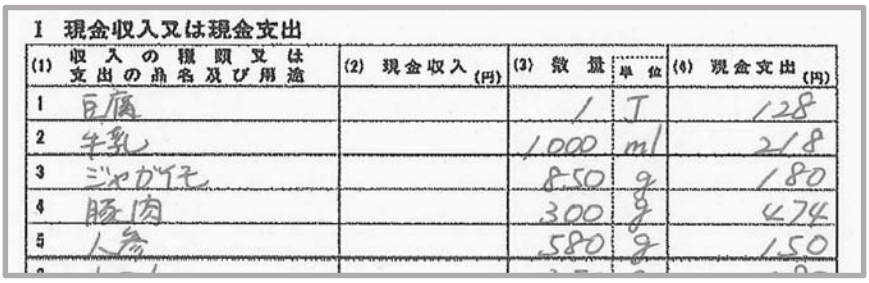
つまり、ある特徴 F_j がレコードに出現し、符号の候補がいくつかあるとき、式 (11) に示すように、その中で n_{kj} が最大となるものが、その特徴に関して最大の事後確率を持つ候補符号となる。次に、式 (12) にあるように、特徴別の最有力候補の中から、式 (10) で定義される $\hat{p}_{kj}$ が最大のものが、レコードに付与する符号となり、併せて出力する信頼度は、擬似信頼度を使い式 (13) に基づき算出しており、Tsubaki et al. (2017) のナイーブベイズによる分類器の事後確率である式 (8) よりも計算が簡単である。

4.　実データによる適用事例とアルゴリズム

4.1　家計調査データについて

ここでは、使用データのソースである家計調査と、分類対象となる収支項目分類について解説する。家計調査は、学生の単身世帯などを除く全国の全ての世帯を対象として、毎月約 9 千世帯に、日々の家計上の収入及び支出を家計簿により調査し、収支の金額や一部数量等の結果を公表している月次の基幹統計（公的統計の根幹をなす重要性の高い統計）である。従来、分類符号の格付を専門とする職員が冊子の形になっている家計簿の記入内容を見て、対応する分類符号を判断し、金額とともにデータ入力することにより、調査票の電子化が行われている。実際の調査票の記入例を、図 1 に示す。例えば、先頭にある「豆腐」の場合、別添に示す家計調査の収支項目分類符号表に従い、対応する三桁の分類符号「280」が入力される。

家計調査では、主に現金の収支を主として、支出の中の消費支出は、さらに商品の種類により同一商品を同じ品目に分類する「品目分類」と、調査対象世帯が購入した商品をその世帯で消費するのか他の世帯に贈るのかという使用目的により分類する「用途分類」の二つの方法により分類され、収支項目分類と総称されるその三桁の符号の総数は約 550 ある。



I 現金収入又は現金支出

(1) 収入の種類又は支出の品名及び用途	(2) 現金収入 (円)	(3) 数量	単位	(4) 現金支出 (円)
1 豆腐		1	T	128
2 牛乳		1,000	ml	218
3 ジャガイモ		850	g	180
4 豚肉		300	g	474
5 人参		580	g	150

図 1. 家計簿の記入例

4.2　システムの概要

提案する自動格付システムは、図 2 に示すように、二つのプロセスに分かれている。プロセス 1 では、学習用データとして、表 1 に示すような三桁の正解符号が付与された家計簿の記入欄の文字列データを使用し、オープンソースの形態素解析エンジン MeCab（Kudo et al., 2004）を用いて、文字列データを単語に分割し、分割された単語から特徴を抽出する自然言語処理を行い、抽出した特徴と対応する正解符号の組合せについて度数分布表を作成する。そして、プロセス 2 では、プロセス 1 と同様に MeCab により文字列を単語に分割して特徴を抽出し、その特徴に対応するラベルと頻度の情報を度数分布表から特定し、確率の計算により最も有力な符号を選択し、信頼度を算出して出力する。

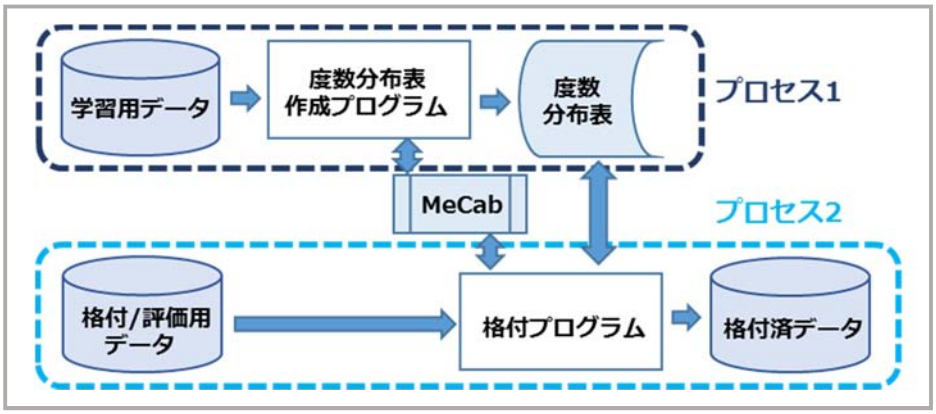


図 2. システムの構成

なお、自然言語処理のためのシステム開発環境として、Windows 上に Cygwin により擬似的な UNIX 環境を構築し、使用・作成するデータの文字コードは UTF-8 に統一した。プログラミング言語には、配列と同様に任意個のデータを格納して自由に取り出すことのできるハッシュというデータ構造を持ち、テキスト処理に優れる Perl を採用した。配列の場合、個々の構成要素を特定するためにインデックスと呼ばれる数値を用いるが、ハッシュの場合は数値ではなく任意のユニークな文字列を使用し、その文字列自体も値として使用することができるという特徴を持つ。以下に、各プロセスの仕組みについて述べ、さらに具体例によりプロセス 2 の具体的な処理の流れを解説する。

表1. 学習用データのイメージ

符号	文字列	符号	文字列
961	おこづかい（3月分　長女）	129	ドーナツ
198	あじの開き	376	惣菜（白和え）
301	みかん	264	トマト
224	牛すじ	278	昆布
368	カキフライ	241	ほうれん草
706	うがい薬	889	コピー代
010	定期収入	102	米
763	携帯代（2台）	260	サヤエンドウ
254	人参	290	しらたき
129	ちくわパン	769	宅急便代（長男）
181	刺身（ブリ）	077	介護保険（世帯主）６期分
762	固定電話料金（1月分）	763	移動電話料金（1月分）

1レコード

(1)　プロセス1

プロセス1では、学習用データを用いて自然言語処理と度数分布表の作成を行う。自然言語処理はMeCabによる文字列の単語分割と特徴の作成から構成される。学習用データの各レコードは、短い文字列と手作業で付与された教師符号で構成される。短い文字列の内容は、世帯における収入や支出について家計簿に記入されるような内容（商品名、サービス名など）を想定している。

まず、短い文字列をMeCabにより単語に分割する。MeCabは単語を分割する際に、単語毎に品詞情報も付与するので、後の特徴作成時に不要となる「&」、「@」、「¥」等の記号を分割後の一連の単語から除く。特徴の抽出には、単語レベルの *n*-gram（*n*=1,2,3,…）モデルを用いて、原則として任意の文字列から切り出されたユニグラム（unigram: 単語一つを一つの特徴とする）、バイグラム（bigram: 切り出された各単語の隣り合う二つの単語の組み合わせを一つの特徴とする）と、文字列の全文を特徴として採用する。家計調査の文字列情報を用いて、考慮する単語数 n を 1 から増やして実験を行った結果に基づき、ユニグラム、バイグラム、全文を特徴として採用した場合が最も精度が高かったため、本稿ではこれらを特徴として採用する。同様に分類器の精度を向上させるため、入力文字列が1単語のみから構成されているとき、その単語の先頭に"+"記号を付与する。この処理により、同じ単語でも入力文字列が1単語で構成される場合と、複数単語から構成される文字列の一部分である場合とで、度数分布を作成する際に異なる特徴として扱うことができる。

分割された単語を用いて特徴を抽出した後、学習用データの各レコードに含まれる教師符号と特徴の組み合わせの出現頻度を集計し、学習用データ全体について度数分布表を作成する。度数分布表のイメージを表2に示す。

例えば、学習用データのあるレコードの記入内容の文字列「クリーミーチキンポットパイ」で、教師符号が「395（外食、洋食）」から構成されていたとする。この場合、1単語からなる特徴として「クリーミー」・「チキン」・「ポット」・「パイ」、2単語からなる特徴として「クリーミー＋チキン」、「チキン＋ポット」・「ポット＋パイ」、そして全文による特徴として「クリーミー＋チキン＋ポット＋パイ」が抽出され、これら全ての特徴と教師ラベル「395」をペアで組み合わせ、度数分布表に足し上げる。全ての学習用データに対してこの処理を行い、度数分布表を完成して、プロセス1は終了する。

実際には、この度数分布表は非常に巨大で度数に 0 の多い疎なものとなるため、特徴と符号の組み合わせがユニークになる特徴・符号・度数の 3 列の表形式データを Perl のハッシュで保持することにより、システムを効率化している。

表 2　度数分布表のイメージ

特徴	符号					
	符号1	符号2	・・・	符号k	・・・	符号M
特徴1	n_{11}	n_{12}	・・・	n_{1k}	・・・	n_{1M}
特徴2	n_{21}	n_{22}	・・・	n_{2k}	・・・	n_{2M}
・・・	・・・	・・・	・・・	・・・	・・・	・・・
特徴 j	n_{j1}	n_{j2}	・・・	n_{jk}	・・・	n_{jM}
・・・	・・・	・・・	・・・	・・・	・・・	・・・
特徴 J	n_{J1}	n_{J2}	・・・	n_{Jk}	・・・	n_{JM}

(2)　プロセス 2

プロセス 2 では、プロセス 1 と同様に、まず入力データの文字列に対して自然言語処理を行い、特徴を抽出する。次に、レコード別に抽出された特徴に付与される可能性のある候補符号を度数分布表から抜き出し、その候補の中から最も確率が高いと思われる符号を選択するとともに、その符号の信頼度を算出する。

入力データの各レコードは、表 1 に示す学習用データの文字列と符号のうちの文字列のみを持つ。性能評価を行う場合には、人手で付与した教師符号を持つデータも使用するが、格付プロセス内でデータの持つ教師符号の情報は使用されない。

まず、任意の入力レコード i から得られた特徴の集合 $J(i)$ について、擬似信頼度 $p_i(\cdot)$ を式 (11) に基づき算出するが、ここで、α と β の値は、Nelder–Mead 法（Nelder and Mead, 1965）により家計調査データについての最適値を探し、それぞれ -0.111111 と -0.444444 を採用する。

次に、特徴毎に得られた候補符号から、式 (12) の条件を満たすものを最終的に付与する符号k_i として選び、この符号の信頼度 v_i を、式 (13) 基づき算出する。このとき、調整用パラメータγ は、経験的に適当と思われる値を複数試した結果に基づき、本稿では $\gamma = 8.5$ を採用する。

(3) プロセス 2 の処理の流れ

入力データの文字列が「アップルパイ」である場合を例に、具体的なプロセス 2 の処理の流れを示す。

① 入力文字列「アップルパイ」に対して自然言語処理を行い、「アップル」、「パイ」、「アップル＋パイ」という 3 つの特徴を抽出する。

② 学習プロセスで作成した表 2 に示す大きな度数分布表から、抽出された 3 つの特徴と合致する特徴の行のみを取り出す。このとき、全ての特徴が度数 0 の列を省き、例えば表 3

のような候補符号と度数分布が得られたものとする。

表３：　度数分布表から抜き出したデータの例

特徴	候補符号（括弧内は符号の説明）				
	300 (りんご)	344 (ケーキ)	349 (キャンデー)	352 (チョコレート)	387 (炭酸飲料)
アップル	543	50	20	3	30
パイ	0	300	0	0	0
アップル+パイ	0	34	0	0	0

③　特徴毎に、最大頻度となる候補符号が特徴別の最有力候補となるので、対応する擬似信頼度を式（10）に基づき計算する。例えば、特徴「アップル」については、最大度数 543 を持つラベル「300（りんご）」が最有力候補となり、その擬似信頼度は、式(10)に基づいて次のように算出される。

$$p_{アップルパイ}(300) = \frac{n_{アップル300} - 0.444444}{n_{アップル} - 0.111111} = \frac{543 - 0.444444}{646 - 0.111111} \approx 0.840$$

同様に特徴「パイ」と「アップル＋パイ」の擬似信頼度は、それぞれ $p_{アップルパイ}(344) \approx 0.999$ と $p_{アップルパイ}(344) \approx 0.990$ になる。特徴「パイ」と「アップル＋パイ」ともラベル「344（ケーキ）」を支持している。このように同じ候補符号について擬似信頼度が 2 つ存在する場合は値が大きい $p_{アップルパイ}(344) \approx 0.999$ を採用する。

④　最終候補として残った、$p_{アップルパイ}(300) \approx 0.840$ と $p_{アップルパイ}(344) \approx 0.999$ から、式（12）に基づき最も擬似信頼度が大きい $p_{アップルパイ}(344) \approx 0.999$ に対応する符号 344 を採用し、この符号に対応する信頼度を式（13）により算出する。

$$v_{アップルパイ} = p_{アップルパイ}(344) \times \frac{p_{アップルパイ}(344)^{8.5}}{p_{アップルパイ}(300)^{8.5} + p_{アップルパイ}(344)^{8.5}}$$

$$\approx 0.999 \times \frac{0.999^{8.5}}{0.840^{8.5} + 0.999^{8.5}} \approx 0.813$$

5.　性能評価

5.1　使用データとその特徴について

まず、学習用データと評価用データを作成する。約 29,000 世帯が含まれる、約 521 万レコード（約 200MB）の家計調査のデータセットを使用し、無作為抽出した 2,327 世帯分のデータ（約 65 万レコード）を評価用データ、残りの約 455 万レコードを学習用データとする。また、精度の安定性を評価するため、無作為抽出による学習用・評価用データの作成と、分類器による格付を 10 回行う。

家計調査のデータセットに含まれるレコードは、表 1 に示したように、世帯が家計簿に記入した、購入商品名や利用したサービス名、収入項目など、世帯の家計に関する項目を記述する短い文字列の情報を含み、次のような特徴がある。

- 文字列情報は、少ない単語数で構成されているものが多く、図 3 に示すように、全体の 98%

のデータは5単語以下で構成されている。

- 使用したデータには、誤った教師符号が付与されているレコードも存在する。
- 文字列情報の中には、方言や表記ゆれ、同音異義語など、熟練した職員でさえ符号付与が難しいと思われるものも含まれている。
- 文字列情報の中には、季節商品や特定の時期に支払う税金など、特定の時期に依存するものも含まれている。
- 文字列情報の内容のみから分類符号を判断することが困難であり、金額情報や世帯の家族構成、家計簿の前後の記入情報など、入力として使用する文字列情報以外の情報を必要とするレコードも存在する。例えば、文字列情報が「シャツ」である場合、分類符号は子ども用、紳士用、婦人用で異なるが、それをこの文字列情報のみから判断することはできない。また、「紅茶」と記載されている場合に、分類符号は茶葉と紅茶飲料で異なるが、それをこの文字列情報のみから判断することは難しい。

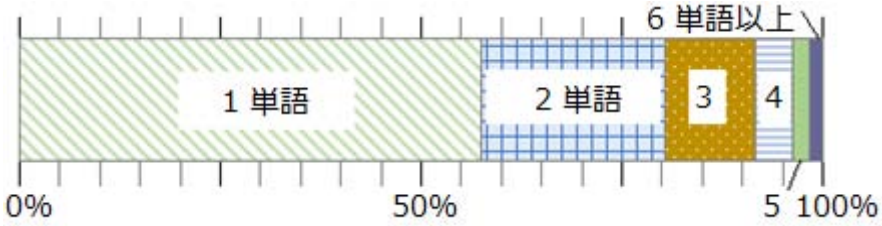


図3. 文字列情報を構成する単語数の分布（構成比）

5.2 家計調査データへの格付システム適用上の工夫

事前分析の結果、収支項目符号には、記入内容が同一でも、家計簿の収入欄に記入される場合と支出欄に記入される場合で符号が異なるものが一定数存在することが判明した。例えば、「預貯金」という文字列情報について、「預貯金の受取」か「預貯金の預入」により付与すべき符号は異なる。格付精度の向上を図るため、収入に関するデータと支出に関するデータを分割し、それぞれプロセス1とプロセス2を別々に実行した。

5.3 評価方法と検証結果

分類器の精度を測定するために、符号付与率（coverage）と正解率（accuracy）を以下のように定義する。

$$\text{符号付与率} = \frac{\text{評価用データのうち符号が付与されたレコード数}}{\text{評価用データのレコード数}}$$

$$\text{正解率} = \frac{\text{評価用データのうち正解符号が付与されたレコード数}}{\text{評価用データのうち符号が付与されたレコード数}}$$

図 4 に、10 回分の検証から得られた符号付与率と正解率の関係を示す。これは、一般的に Precision-Recall 曲線と呼ばれるものを利用した評価方法であり、適合率（Precision）と再現率（Recall）を用いて評価を行う。適合率とは、正確性に関する指標で、システムが出力した結果に対して、真に正しかったものの割合のことをいい、本稿においては正解率に相当する。一方で再現

率とは、網羅性に関する指標で、結果として出力されるべきもののうち、実際にシステムが出力したものの割合のことをいい、本稿においては符号付与率に相当する。一般的に、適合率と再現率はトレードオフの関係性を持ち、図４で示した符号付与率と正解率の関係についても同様で、正解率の閾値を 99.5%に設定した場合に、評価用データの約 50%に符号を付与することができ、正解率の閾値を 98%とする場合は、評価用データの約 78%に符号を付与することができる。また、この符号付与率と正解率のカーブは、検証回数分描かれているが、これらの線はほぼ重なっており、これは性能が安定していることを示している。そして、提案する分類器が符号を付与したデータ全体の正解率は、約 90.8%である。

一般的に人手でデータに符号を付与した場合の正解率が、おおよそ 98%程度と言われているため、閾値として特に 98%の正解率に着目し、この条件での符号付与率を表４に示した。

また、全体の精度の評価に加えて、実用化に向けてより詳細な分析を行うために、部門別の分析を行い、その結果を表５に示した。この表は、約 550 ある項目符号を、より大きな単位となる部門でまとめ、部門ごとに分類符号の正解率を集計しており、部門により正解率が大きく異なることがわかる。正解率が 90%を超えている部門が半数以上を占める一方で、部門Ｄ（被服及び履物）や部門Ｆ（教育）については正解率が低い。不正解の内容を分析した結果、部門Ｄは、例えば 4.2 節で述べたように、被服や履物などは誰のものかで符号が異なる場合が多く、同様に部門Ｆも、単に子供の学費であっても、それが私立なのか公立なのか等で分類符号が異なり、符号格付には文字列情報以外の情報も必要となることが主たる原因である。

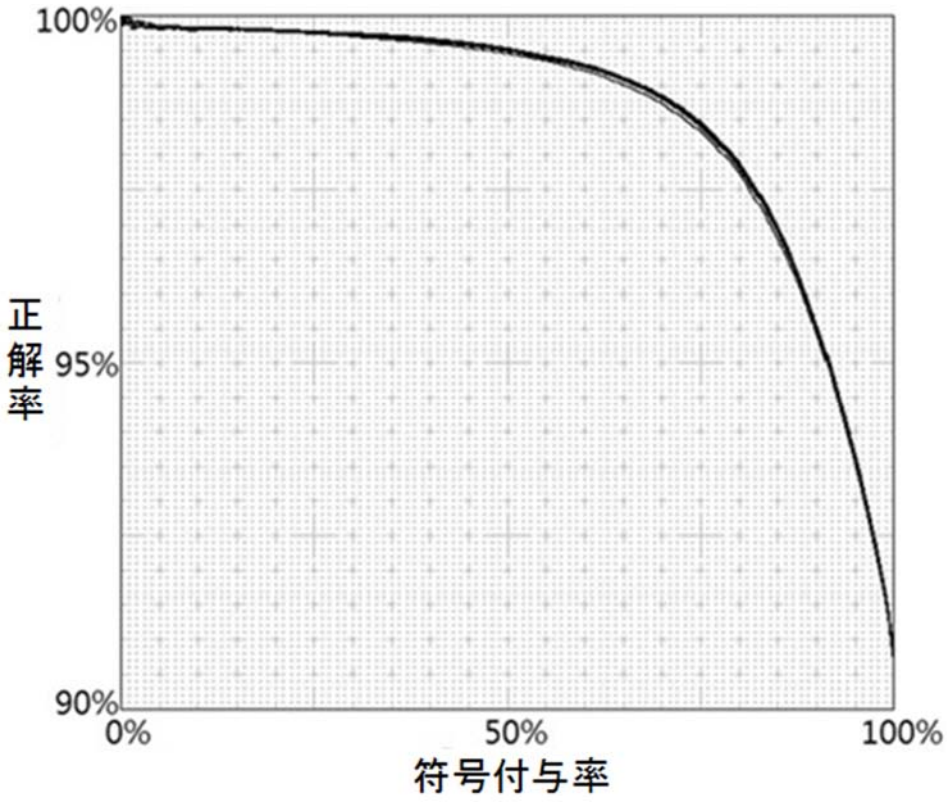


図 4. 正解率と符号付与率の関係

表 4.　正解率 98%を条件とした場合の結果一覧

試行 No.	対象 レコード数 (1)	符号 付与率 (2) / (1)	正解率 (3) / (2)	符号 付与数 (2)	正解数 (3)
1	655,573	78.81%	98.00004%	516,661	506,328
2	654,991	78.89%	98.00018%	516,696	506,363
3	648,257	78.71%	98.00006%	510,266	500,061
4	653,043	77.95%	98.00009%	509,074	498,893
5	652,376	78.59%	98.00018%	512,695	502,442
6	648,609	79.19%	98.00012%	513,631	503,359
7	657,003	78.98%	98.00002%	518,906	508,528
8	648,991	78.86%	98.00001%	511,802	501,566
9	652,120	79.19%	98.00007%	516,419	506,091
10	655,161	78.84%	98.00011%	516,528	506,198

表 5.　部門別の正解率

部門	部門名	正解	不正解	正解率
A	実収入,実収入以外の受け取り	121,792	11,589	91.31%
B	食料	4,523,027	394,615	91.98%
C	住居、光熱･水道	72,267	5,992	92.34%
D	被服及び履物	77,664	28,162	73.39%
E	保健医療、交通･通信	277,710	28,833	90.59%
F	教育	7,566	6,519	53.72%
G	家具･家事用品	643,214	109,422	85.46%
H	教養娯楽、非消費支出,実支出以外の支払	174,599	14,170	92.49%
I	その他	25,172	3,811	86.85%

6.　他手法との比較

ここでは、本システムの開発以前に導入が検討されていた、家計調査の収支項目分類の格付のためにチューニングされた商用のディープラーニングによる分類器を用いた試算結果と比較をするため、可能な限り条件を揃え、提案手法で試算を行った。

第 5 節と同様に、家計調査のデータセットから無作為抽出し、約 455 万レコードを学習用データ、約 65 万レコードを評価用データとして使用した。結果は図 5 のとおりで、ディープラーニングによる分類器の結果を▽印、提案する分類器による結果を○印で示している。ディープラーニングによる分類器の精度は、図 5 に矢印で示す正解率 98%の閾値に着目した場合の符号付与率では、本稿で提案する分類器よりも精度が良いが、符号付与率と正解率との関係性を示すカーブを全体的にみる限り、精度に大きな差異はないといえる。

次に、この二つの分類器の処理時間を表 6 にまとめた。評価には、家計簿のデータセットから 50

万レコードを使用し、それぞれの分類器について学習プロセスと評価プロセスの処理時間を測定している。ただし、提案手法については、プロセス 2 の一部に、学習プロセスの一部も含まれるが、システム上切り分けることが難しいため、便宜上、プロセス 1 を学習プロセス、プロセス 2 を評価プロセスとして、処理時間の比較を行っている。ディープラーニングによる分類器は並列計算を行うが、提案する分類器は並列計算の機能を持たず、さらに実験を行った計算環境も同じではないが、学習プロセスも評価プロセスについても、提案する分類器の処理速度が圧倒的に速いという結果が得られた。

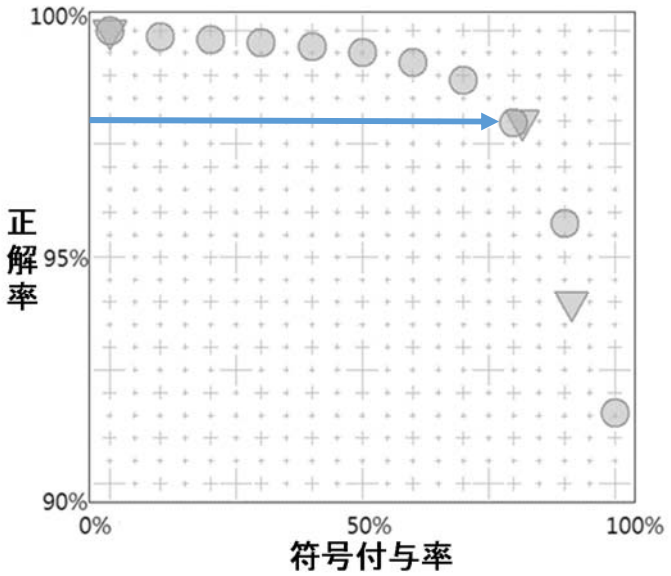


図 5.　比較結果

表 6. 処理時間

	提案手法	ディープラーニングアルゴリズム
設定	Xeon, 2.3 GHz, 並列処理なし	8 コア, 8 並列処理
学習プロセス処理時間	32.3 秒	6 時間
評価プロセス処理時間	56.5 秒	2 時間

7.　結論と課題

提案する自動格付システムは、数百もの符号を持つ多クラス分類を、高速に、商用システムに準じる高い精度で行うことができる。また、汎用の分類器なので家計調査以外の格付業務にも幅広く適用できる可能性を持ち、利用にあたりライセンス料なども発生しない。

このシステムの特徴として、符号を予測するだけではなく「信頼度」と呼ぶ付与した符号の確からしさを併せて算出することができるが、これは実用上とても重要な機能である。例えば、格付処理の目標精度を 98%と設定する場合、何らかのチェックが必要になるのは、信頼度が 98%未満のレコードということになり、全データの約 78%を占める信頼度 98%以上のデータを自動格付し、統計作成のための次の行程に人手を介すことなく送ることができるということになる。

家計調査のデータセットを用いた評価の結果、提案する分類器は、大部分のデータに対して、正しい符号を安定的に付与できるといえる。ただし、部門別の結果に見られるように、個々の符号別にみれば、必ずしも正解率が高くないもの存在するため、今後の実用化に向けて、さらなる性能向上の取り組みが必要である。

また、もう一つの課題は、どのくらいの学習用データが必要なのかという検討である。第 5.1 節で使用した約 521 万レコードが、現在利用可能な家計調査データの全てであり、そのうちの約 455 万レコードの学習用データを 50 万レコードまで徐々に減らして符号付与を行った結果を表 7 に示す。学習用データのレコード数が大きいほど、改善の割合は逓減するが、符号付与率及び正解率は改善されていることがわかる。このため、今後さらにデータの蓄積が進んだ段階で、必要なデータ量について検討する必要がある。

さらに、約 521 万の全レコードを符号別にみると、表 8 のとおり、約 550 の符号数のうち、70（70＝13＋57）種類の符号の出現回数がテータセット全体でも 100 未満であることがわかる。学習用データ量の検討については、このような出現頻度の低い符号も考慮する必要がある。

表 7. 学習データのレコード数別の結果

学習データ レコード数	対象 レコード数	符号 付与数	正解数	符号 付与率	正解率
500,000	655,573	643,066	565,826	98.09%	87.99%
1,000,000	655,573	647,730	577,869	98.80%	89.21%
2,000,000	655,573	650,289	585,525	99.19%	90.04%
3,000,000	655,573	651,264	588,784	99.34%	90.41%
4,000,000	655,573	651,819	591,053	99.43%	90.68%

表 8. 家計調査のデータセットにおける符号の出現回数の度数分布表

度数	分類符号の数
1-9	13
10-99	57
100-999	109
1,000-9,999	259
10,000-99,999	146
100,000-	2

8. 考察

この研究は、平成 26 年度から現在までの、家計調査の機械学習アルゴリズムによる自動格付の取り組みの一部であり、ディープラーニングに代表されるブラックボックス型のシステムを、公的統計の作成に使用することの是非も課題の一つであった。近年急速に普及した機械学習のアルゴリズムは、従来の統計手法では難しい課題にも取り組むことができるが、一方で、アウトプットがなぜそのような結果になったかを説明するのは難しく、開発だけでなく課題に合わせるためのチューニングの難易度も高いものが多い。そのような複雑なアルゴリズムは、既存のフリーのソフトウェアを利用すれば内部開発ができるかもしれないが、継続的な運用には困難が予想される。このため、家計調査データの収支項目分類の格付という課題解決を目指し、まず平成 26 年度に商用システム

の導入を検討するため、第 6 節で述べたディープラーニングによるシステムを用いた委託研究が実施され、良い結果が得られた。平成 27 年度からは、この商用システムの費用対効果の検証を兼ねて、結果の説明がしやすく内部での開発が可能で、より運用に問題が少ないと思われる、比較的単純な仕組みの、分類器を開発、28 年度末にはこの分類器を用いたシステムを既存のルールベース型のシステムと併用することにより、実用化する方針が示された。

本稿で提案する分類器によるこのシステムは、分類格付の結果について、なぜその符号が付与されたのかを比較的簡単に説明することができ、組織内での開発や継続的な運用も可能な程度に単純な仕組みであるが、家計調査の収支項目符号のような短い文章による分類格付という課題については、専門家のチューニングを施した商用ディープラーニングシステムに近い精度を達成している。

参考文献

[1] 田口玄一 (1997),「品質工学の数理、7. 率のデータ, 特に科学, 生物分野のデータと SN 比」, 品質工学, Vol. 5, No.2, pp. 3-9.

[2] 油井清吾 (2017),「分類符号格付への統計的機械学習の適用」, 統計, 2017 年 1 月号, pp. 14-19. 一般財団法人日本統計協会.

[3] Copas, J. B. (1983), “Regression, Prediction and Shrinkage,” Journal of the Royal Statistical Society, Series B 45, pp. 311-354.

[4] Kudo, T., Yamamoto, K., Matsumoto, Y. (2004), “Applying Conditional Random Fields to Japanese Morphological Analysis,” Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp. 230-237.

[5] Nelder, J. A. and Mead, R. (1965), “A simplex algorithm for function minimization,” The Computer Journal, Vol. 7, Issue 4, pp. 308-313.

[6] Toko, Y., Wada, K., and Kawano, M. (2017), “A supervised multiclass classifier for an Autocoding System,” Journal of Romanian Statistical Review, 4, pp. 29-39.

[7] Tsubaki, H., Wada, K., and Toko, Y. (2017), “An extension of Taguchi's T method and standardized misclassification rate for supervised classification with only binary inputs,” to appear in Proceedings of the ANQ Congress 2017, Kathmandu, Nepal.

＜家計調査 家計収支編 収支項目分類一覧（平成27年（2015年）1月～）＞

別添

010-052	**受 取**
010-039	**実収入**
010-030,033-035	経常収入
010-014	勤め先収入
010-012	世帯主収入
010	定期収入
011	臨時収入
012	賞 与
013	世帯主の配偶者の収入
014	他の世帯員収入
020-022	事業・内職収入
022	家賃収入
020	他の事業収入
021	内職収入
023	農林漁業収入
023	農林漁業収入
030,033-035	他の経常収入
030	財産収入
034,035	社会保障給付
034	公的年金給付
035	他の社会保障給付
033	仕送り金
032,039	特別収入
032	受贈金
039	他の特別収入
040-049,052	**実収入以外の受取（繰入金を除く）**
040	預貯金引出
048,052	保険金
048	個人・企業年金保険金
052	他の保険金
045	有価証券売却
047	土地家屋借入金
042	他の借入金
043	分割払購入借入金
044	一括払購入借入金
046	財産売却
049	実収入以外の受取のその他
050	繰入金
050	繰入金
102-981,070-092	**支払**
102-981,070-079	**実支出**
102-981	**消費支出**
102-399,39A, 39B,39X,39Y	**食 料**
102-160	穀 類
102	米
102	米
120,129	パ ン
120	食パン
129	他のパン
130-139	麺 類
130	生うどん・そば
131	乾うどん・そば
134	スパゲッティ
133	中華麺
135	カップ麺
132	即席麺
139	他の麺類
140-160	他の穀類
140	小麦粉
150	も ち
160	他の穀類のその他
170-217	魚介類
170-194	生鮮魚介
170-189	鮮 魚
170	まぐろ
172	あ じ
173	いわし
174	かつお
175	かれい
176	さ け
177	さ ば
178	さんま
180	た い
181	ぶ り
182	い か
183	た こ
185	え び
186	か に
189	他の鮮魚
187	さしみ盛合わせ
190-194	貝 類
190	あさり
192	しじみ
191	かき（貝）
194	ほたて貝
193	他の貝
195-202	塩干魚介
195	塩さけ
196	たらこ
197	しらす干し
198	干しあじ
202	他の塩干魚介
203-209	魚肉練製品
203	揚げかまぼこ
204	ちくわ
205	かまぼこ
209	他の魚肉練製品
210-217	他の魚介加工品
210	かつお節・削り節
213	魚介の漬物
215	魚介のつくだ煮
216	魚介の缶詰
217	他の魚介加工品のその他
220-229,22X	肉 類
220-224,22X	生鮮肉
220	牛 肉
221	豚 肉
222	鶏 肉
22X	合いびき肉
224	他の生鮮肉
225-229	加工肉
225	ハ ム
226	ソーセージ
227	ベーコン
229	他の加工肉
230-238	乳卵類
230	牛 乳
230	牛 乳
231-235	乳製品
231	粉ミルク
232	ヨーグルト
233	バター
234	チーズ
235	他の乳製品
238	卵
238	卵
240-299,26B,26X	野菜・海藻
240-269,26B,26X	生鮮野菜
240-249	葉茎菜
240	キャベツ
241	ほうれんそう
242	はくさい
243	ね ぎ
244	レタス
247	ブロッコリー
245	もやし
249	他の葉茎菜
250-259,25X	根 菜
250	さつまいも
251	じゃがいも
252	さといも
253	だいこん
254	にんじん
255	ごぼう
256	たまねぎ
258	れんこん
25X	たけのこ
259	他の根菜
260-269,26B,26X	他の野菜
260	さやまめ
261	かぼちゃ
262	きゅうり
263	な す
264	トマト
265	ピーマン
266	生しいたけ
26B	しめじ
26X	えのきたけ
267	他のきのこ
269	他の野菜のその他
273-279	乾物・海藻
273	豆 類
274	干ししいたけ
276	干しのり
277	わかめ
278	こんぶ
279	他の乾物・海藻
280-289	大豆加工品
280	豆 腐
281	油揚げ・がんもどき
282	納 豆
289	他の大豆製品
290-299	他の野菜・海藻加工品
290	こんにゃく
291	梅干し
292	だいこん漬
293	はくさい漬
294	他の野菜の漬物
295	こんぶつくだ煮
296	他の野菜・海藻のつくだ煮
299	他の野菜・海藻加工品のその他
300-319	果 物
300-316	生鮮果物
300	りんご
301	みかん
314	グレープフルーツ
315	オレンジ
304	他の柑きつ類
305	梨
306	ぶどう
307	柿
308	桃
309	すいか
310	メロン
311	いちご
312	バナナ
316	キウイフルーツ
313	他の果物
319	果物加工品
319	果物加工品

総務省統計局HP(http://www.stat.go.jp/data/kakei/koumoku/bunrui27.htm)より取得し、配置を変更した。

320-339,33X	油脂･調味料
320,321	油　脂
320	食用油
321	マーガリン
322-339,33X	調味料
322	食　塩
323	しょう油
324	み　そ
325	砂　糖
327	酢
328	ソース
329	ケチャップ
330	マヨネーズ･マヨネーズ風調味料
332	ドレッシング
331	ジャム
333	カレールウ
334	乾燥スープ
335	風味調味料
336	ふりかけ
33X	つゆ･たれ
339	他の調味料
340-359	菓子類
340	ようかん
341	まんじゅう
342	他の和生菓子
343	カステラ
344	ケーキ
347	ゼリー
348	プリン
345	他の洋生菓子
350	せんべい
346	ビスケット
357	スナック菓子
349	キャンデー
352	チョコレート
353	チョコレート菓子
356	アイスクリーム･シャーベット
359	他の菓子
360-376	調理食品
360-363,36A,36B	主食的調理食品
360	弁　当
36A	すし(弁当)
36B	おにぎり･その他
361	調理パン
363	他の主食的調理食品
364-376	他の調理食品
364	うなぎのかば焼き
365	サラダ
366	コロッケ
367	カツレツ
368	天ぷら･フライ
369	しゅうまい
371	ぎょうざ
372	やきとり
373	ハンバーグ
370	冷凍調理食品
375	そうざい材料セット
376	他の調理食品のその他
380-389,38A,38B,38X,38Y	飲　料
380,381,383,38X	茶　類
380	緑　茶
381	紅　茶
383	他の茶葉
38X	茶飲料
382,384,386	コーヒー･ココア
382	コーヒー
384	コーヒー飲料
386	ココア･ココア飲料
385,387-389,	
38A,38B,38Y	他の飲料
385	果実･野菜ジュース
387	炭酸飲料
388	乳酸菌飲料
38A	乳飲料
38B	ミネラルウォーター
38Y	スポーツドリンク
389	他の飲料のその他
3X1-3X9,3XX	酒　類
3X1	清　酒
3X3	焼　酎
3X4	ビール
3X5	ウイスキー
3X9	ワイン
3X7	発泡酒･ビール風アルコール飲料
3X8	チューハイ･カクテル
3XX	他の酒
390-399,39A,39B,39X	外　食
390-399,39A,39B	一般外食
390-396,399,39A,39B	食事代
390	日本そば･うどん
391	中華そば
392	他の麺類外食
393	すし(外食)
394	和　食
39A	中華食
395	洋　食
399	焼　肉
39B	ハンバーガー
396	他の主食的外食
397	喫茶代
398	飲酒代
39X	学校給食
39X	学校給食

39Y	賄い費
39Y	賄い費
400-429	**住　居**
400-409	家賃地代
400	民営家賃
403	公営家賃
404	給与住宅家賃
402	地　代
409	他の家賃地代
410-429	設備修繕･維持
410,419	設備材料
410	設備器具
419	修繕材料
420-429	工事その他のサービス
420	畳替え
424	給排水関係工事費
425	外壁･塀等工事費
426	植木･庭手入れ代
427	他の工事費
429	火災･地震保険料
430-440	**光熱･水道**
430,43X	電気代
43X	深夜電力電気代
430	他の電気代
431,432	ガス代
431	都市ガス
432	プロパンガス
433,439	他の光熱
433	灯　油
439	他の光熱のその他
440	上下水道料
440	上下水道料
451-542	**家具･家事用品**
451-489	家庭用耐久財
451-459,45X	家事用耐久財
45X	電子レンジ
451	炊事用電気器具
452	炊事用ガス器具
453	電気冷蔵庫
455	電気掃除機
456	電気洗濯機
459	他の家事用耐久財
470-479	冷暖房用器具
470	エアコンディショナ
472	ストーブ･温風ヒーター
479	他の冷暖房用器具
480-489	一般家具
480	たんす
481	食卓セット
482	応接セット
483	食器戸棚
489	他の家具
491-499	室内装備･装飾品
491	照明器具
492	室内装飾品
493	敷　物
496	カーテン
499	他の室内装備品
500-509	寝具類
500	ベッド
501	布　団
503	毛　布
505	敷　布
509	他の寝具類
510-529	家事雑貨
510	茶わん･皿･鉢
514	他の食卓用品
515	鍋･やかん
517	他の台所用品
518	電球･ランプ
519	タオル
529	他の家事雑貨
530-539	家事用消耗品
531,532	ティッシュペーパー･トイレットペーパー
531	ティッシュペーパー
532	トイレットペーパー
533,534	洗剤
533	台所･住居用洗剤
534	洗濯用洗剤
530,535-539	他の家事用消耗品
530	ポリ袋･ラップ
535	殺虫･防虫剤
536	柔軟仕上剤
537	芳香･消臭剤
539	他の家事用消耗品その他
540-542	家事サービス
540	家事代行料
541	清掃代
542	家具･家事用品関連サービス
550-694	**被服及び履物**
550-558	和　服
550	男子用和服
552	婦人用着物
554	婦人用帯
557	他の婦人用和服
558	子供用和服
560-582	洋　服
560-569	男子用洋服
560	背広服
561	男子用上着
562	男子用ズボン
563	男子用コート
565	男子用学校制服
569	他の男子用洋服

総務省統計局HP(http://www.stat.go.jp/data/kakei/koumoku/bunrui27.htm)より取得し、配置を変更した。

570-576	婦人用洋服
570	婦人服
574	婦人用上着
571	スカート
572	婦人用スラックス
573	婦人用コート
575	女子用学校制服
576	他の婦人用洋服
580,582	子供用洋服
580	子供服
582	乳児服
590-597	シャツ･セーター類
590-592	男子用シャツ･セーター類
590	ワイシャツ
591	他の男子用シャツ
592	男子用セーター
593-595	婦人用シャツ･セーター類
593	ブラウス
594	他の婦人用シャツ
595	婦人用セーター
596,597	子供用シャツ･セーター類
596	子供用シャツ
597	子供用セーター
600-621	下着類
600,602	男子用下着類
600	男子用下着
602	男子用寝巻き
610-614	婦人用下着類
610	婦人用ファンデーション
612	他の婦人用下着
614	婦人用寝巻き
620,621	子供用下着類
620	子供用下着
621	子供用寝巻き
631,640	生地･糸類
631	着尺地･生地
640	他の生地･糸類
650-659	他の被服
650	帽　子
651	ネクタイ
652	マフラー･スカーフ
653	手袋
654	男子用靴下
655	婦人用ストッキング
656	婦人用ソックス
657	子供用靴下
659	他の被服のその他
670-680	履物類
675	運動靴
679	サンダル
670	男子靴
672	婦人靴
676	子供靴
680	他の履物
691-694	被服関連サービス
691	洗濯代
694	被服賃借料
692	他の被服関連サービス
700-729	**保健医療**
700-709	医薬品
700	感冒薬
701	胃腸薬
702	栄養剤
704	外傷･皮膚病薬
706	他の外用薬
709	他の医薬品
710	健康保持用摂取品
710	健康保持用摂取品
711-719	保健医療用品･器具
713	紙おむつ
711	保健用消耗品
712	眼　鏡
714	コンタクトレンズ
719	他の保健医療用品･器具
720-729	保健医療サービス
720	医科診療代
722	歯科診療代
723	出産入院料
721	他の入院料
724	整骨（接骨）･鍼灸院治療代
728	マッサージ料金等（診療外）
727	人間ドック等受診料
729	他の保健医療サービス
730-769	**交通･通信**
730-739	交　通
730	鉄道運賃
731	鉄道通学定期代
732	鉄道通勤定期代
733	バス代
734	バス通学定期代
735	バス通勤定期代
736	タクシー代
737	航空運賃
738	有料道路料
739	他の交通
740-759,75B,75X	自動車等関係費
740,742	自動車等購入
740	自動車購入
742	自動車以外の輸送機器購入
745	自転車購入
745	自転車購入
750-759,75B,75X	自動車等維持
750	ガソリン
751	自動車等部品
752	自動車等関連用品
753	自動車整備費
755	自動車以外の輸送機器整備費
75X	年極･月極駐車場借料
756	他の駐車場借料
75B	レンタカー･カーシェアリング料金
754	他の自動車等関連サービス
757	自動車保険料(自賠責)
758	自動車保険料(任意)
759	自動車保険料以外の輸送機器保険料
760-769	通　信
760	郵便料
762	固定電話通信料
763	移動電話通信料
769	運送料
766	移動電話
764	他の通信機器
770-792	**教　育**
770-779	授業料等
770	国公立小学校
771	私立小学校
772	国公立中学校
773	私立中学校
774	国公立高校
775	私立高校
776	国公立大学
777	私立大学
778	幼児教育費用
779	専修学校
780,781	教科書･学習参考教材
780	教科書
781	学習参考教材
790-792	補習教育
790	幼児･小学校補習教育
791	中学校補習教育
792	高校補習教育･予備校
801-889,88A,88B,88X,88Y	**教養娯楽**
801-813	教養娯楽用耐久財
801	テレビ
803	携帯型音楽･映像用機器
813	ビデオレコーダー･プレイヤー
810	パーソナルコンピュータ
804	カメラ
811	ビデオカメラ
806	楽　器
807	書斎･学習用机･椅子
809	他の教養娯楽用耐久財
812	教養娯楽用耐久財修理代
821-849,84A,84X,84Y	教養娯楽用品
821-829	文房具
821	筆記･絵画用具
826	ノート･紙製品
827	他の学習用消耗品
828	他の学習用文房具
829	他の文房具
832-834	運動用具類
832	ゴルフ用具
833	他の運動用具
834	スポーツ用品
835-837	玩　具
836	テレビゲーム機
835	ゲームソフト等
837	他の玩具
840	切り花
842,843,845-849, 84A,84Y	他の教養娯楽用品
846	音楽･映像用未使用メディア
845	音楽･映像収録済メディア
848	ペットフード
84Y	他の愛玩動物･同用品
84A	園芸用植物
847	園芸用品
843	手芸･工芸材料
849	電　池
842	他の教養娯楽用品のその他
84X	動物病院代
841	他の愛玩動物関連サービス
844	教養娯楽用品修理代
850-859	書籍･他の印刷物
850	新　聞
851	雑誌（週刊誌を含む）
854	書　籍
859	他の印刷物
860-889,88A,88B,88X,88Y	教養娯楽サービス
860	宿泊料
860	宿泊料
861,862	パック旅行費
861	国内パック旅行費
862	外国パック旅行費
870-876,879	月謝類
875	語学月謝
870	他の教育的月謝
876	音楽月謝
871	他の教養的月謝
872	スポーツ月謝
873	自動車教習料
874	家事月謝
879	他の月謝類
877,878,880-889,88A,88B,88X,88Y	他の教養娯楽サービス
880,88A,88B	放送受信料
88A	NHK放送受信料
88B	ケーブルテレビ放送受信料
880	他の放送受信料
877,878,881-886	入場･観覧･ゲーム代
882	映画･演劇等入場料
883	スポーツ観覧料
877	ゴルフプレー料金
878	スポーツクラブ使用料
881	他のスポーツ施設使用料

総務省統計局HP(http://www.stat.go.jp/data/kakei/koumoku/bunrui27.htm)より取得し、配置を変更した。

884	文化施設入場料
886	遊園地入場･乗物代
885	他の入場･ゲーム代
888	諸会費
887	写真撮影･プリント代
88X	教養娯楽賃借料
88Y	インターネット接続料
889	他の教養娯楽サービスのその他
890-981	**その他の消費支出**
890-959,95X	諸雑費
890-899	理美容サービス
890	温泉･銭湯入浴料
891	理髪料
892	パーマネント代
894	カット代
899	他の理美容代
900-915	理美容用品
900	理美容用電気器具
901	歯ブラシ
903	他の理美容用品
904-915	石けん類･化粧品
904	浴用･洗顔石けん
905	シャンプー
908	ヘアリンス･ヘアトリートメント
906	歯磨き
907	整髪･養毛剤
909	化粧クリーム
910	化粧水
914	乳液
911	ファンデーション
912	口　紅
915	ヘアカラーリング剤
913	他の化粧品
920-935	身の回り用品
920	傘
920	傘
924-927	かばん類
924	ハンドバッグ
925	通学用かばん
926	旅行用かばん
927	他のバッグ
928	装身具
930	腕時計
932	他の身の回り用品
935	身の回り用品関連サービス
940	たばこ
940	たばこ
950-959,95X	他の諸雑費
950	信仰･祭祀費
955	祭具･墓石
956	婚礼関係費
957	葬儀関係費
958	他の冠婚葬祭費
95X	医療保険料
952	他の非貯蓄型保険料
953	寄付金
954	保育費用
951	介護サービス
959	他の諸雑費のその他
960,961	こづかい(使途不明)
960	世帯主こづかい （単身世帯は「使途不明金」）
961	他のこづかい
970-973	交際費
970	贈与金
970	贈与金
971-973	他の交際費
971	つきあい費
973	住宅関係負担費
972	他の負担費
980,981	仕送り金
980	国内遊学仕送り金
981	他の仕送り金
070-079	**非消費支出**
070,071,075	直接税
070	勤労所得税
075	個人住民税
071	他の税
073,074,076,077	社会保険料
073	公的年金保険料
074	健康保険料
077	介護保険料
076	他の社会保険料
079	他の非消費支出
080-089,092	**実支出以外の支払（繰越金を除く）**
080	預貯金
083,092	保険料
083	個人･企業年金保険料
092	他の保険料
086	有価証券購入
088	土地家屋借金返済
082	他の借金返済
084	分割払購入借入金返済
085	一括払購入借入金返済
087	財産購入
089	実支出以外の支払のその他
090	**繰越金**
090	繰越金

総務省統計局HP(http://www.stat.go.jp/data/kakei/koumoku/bunrui27.htm)より取得し、配置を変更した。