修 士 論 文

ニューラルネットワークを用いた有効な学習素性の選択

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

伊藤　謙

2014 年 9 月

修 士 論 文

ニューラルネットワークを用いた有効な学習素性の選択

指導教員　白井　清昭 准教授

審査委員主査　白井　清昭 准教授
審査委員　飯田　弘之 教授
審査委員　池田　心 准教授

北陸先端科学技術大学院大学
情報科学研究科情報科学専攻

1110008 伊藤　謙

提出年月: 2014 年 8 月

概 要

近年、多くの自然言語処理に機械学習の技術が使われている。機械学習でしばしば問題となるのが過学習である。過学習とは、必要以上の素性を用いることで、訓練データに特化しすぎたモデルが学習されることである。この問題を解決するために取り組まれている手法に素性選択がある。素性選択とは、与えられた素性の集合から有効な素性をみつける手法である。本研究では、素性選択にニューラルネットワーク (NN) を用いた手法を提案する。提案手法では、三層からなる NN を学習する。NN 内のノード間の重みに着目して素性の有効性を評価し、評価の高い素性を選択した素性集合を作成する。また、サポートベクターマシン (SVM) は多くの自然言語処理で良好な成果が得られている機械学習アルゴリズムであることから、素性選択後の素性集合を用いて SVM を学習し、最終的な分類モデルを得る。素性には、ユニグラム、バイグラム、共起単語を使用する。本提案手法では、素性のスコアとして $score_A$、$score_B$、$score_C$ の 3 種類のモデルを提案する。$score_A$ は、素性 i に対応する入力層のノードと中間層のノードとのリンクの重みのみから求めている。$score_B$ は、入力層と中間層間のノードのリンクの重みに加え、中間層と出力層間のノードのリンクの重みも利用している。$score_C$ は、$score_B$ のように入力層と中間層、中間層と出力層の間のリンクの重みを利用しているが、後者については 2 つの出力ノードのうち重みが高い方の重みのみを用いるモデルである。提案手法による素性選択手法を評価するために、テキスト分類をタスクとする実験を行った。最初に、素性の有効性を比較するため、ユニグラム、ユニグラムとバイグラム、ユニグラムと共起単語を素性とし、頻度によって素性選択する 3 つのベースラインを比較した。テキスト分類の F 値が高いのはユニグラム、ユニグラムと共起単語、ユニグラムとバイグラムの順で、それぞれ 0.838、0.742、0.734 であった。次に、提案手法の 3 種類のスコアによって素性選択したモデルを比較したところ、スコアの違いによって大きな差はなかったが、$score_B$(素性がユニグラムとバイグラム、素性数が 15000 のとき) の F 値が一番高かった。提案手法とベースラインの比較では、ユニグラムと共起単語を素性としたときのみ提案手法はベースラインよりも上回った。具体的には、提案手法の $score_C$(素性数が 20000 のとき) の F 値は 0.834 で、ベースラインの 0.742 より 0.092 高かった。ただし、ユニグラムと共起単語を素性とした提案手法の F 値は、ユニグラムのみを素性としたベースラインと同等であった。最後に、提案手法によるスコア付けの妥当性を検証するために、提案手法ならびに出現頻度の降順に素性を減らしたときの F 値の変動を調べた。もし提案手法のスコアが素性の有効性を測る指標として妥当なら、素性を減らすごとに評価指標の値は単調に減少すると考えられる。出現頻度の順に素性を削除したとき、F 値は向上したり低下したりしたが、提案手法のスコアの順に素性を削除したとき、F 値は概ね単調に減少した。この結果から、提案手法のスコア付けは妥当と考えられる。

目 次

図 目 次

表 目 次

第1章　序論

本章では、本研究の背景と研究目的について述べる。

1.1　研究の背景

近年、IT 技術の発展により、企業や官公庁など様々な組織において電子データの保存・管理が進んでいる。また、スマートフォンなどのモバイル端末の普及に伴う利便性の向上から、ウェブには年々膨大な量の情報が蓄積されている。このため、膨大なテキストデータの中から効率的に有用な情報を抽出することは、重要な問題として認識されている。この問題の解決のため、自然言語処理分野でも多くの研究が盛んに行われている。特に、機械学習を用いて自然言語処理における様々な問題を解決する手法が研究されている。

機械学習は、多数のサンプルデータから、そのデータにみられるパターンなどを学習し、その学習結果から未知のデータを正しく分類する技術である。機械学習を行う時、学習データから最適な分類モデルを構築するための重要な要素の 1 つが素性である。素性は、データの内容を表現する情報であり、学習や未知データの分類の際、入力として与えられる情報のことでもある。一般に問題を解決するために有効な手がかりとなる情報が素性となる。したがって、データをどのような素性を用いて表現するかは重要である。例えば、明日の天気について予測するモデルを機械学習で獲得する際、「気温」「気圧」「他の場所の天気」などを素性として用いる。

しかし、機械学習を行う際、あまりに多くの素性を使用すると以下のような問題が生じる。

- 学習に要する計算コスト
- 多数の素性の使用による学習への悪影響

1 つめは、たとえ 1 つの素性に対する計算時間が少なくとも、1 万個以上と多くの素性を使用することにより、多くの計算時間とメモリを学習時に要するという問題である。2 つめは、素性として多くの情報を使用しても、その中には分類に有効なものとそうでないものがあり、有効でない素性が使われると機械学習がうまくいかなくなるという問題である。また、あまりに多くの素性を使用することは過学習を引き起こす危険性が高まる。過学習とは、サンプルデータに対して特化しすぎた分類モデルが学習されてしまうという問題である。過学習された分類モデルは未知のデータに対する分類を誤りやすい。

上記の問題を解決する手法として取り組まれているのが素性選択という手法である。素性選択とは、与えられた素性の集合から、学習に有効な素性の部分集合をみつけるものである。学習に有効な部分集合の見つけ方としては、ある尺度を基に素性の評価を行い、その評価結果から有効と思われる素性の選別を行う手法などがある。素性選択に関する多くの先行研究では、学習データの情報などを基に素性の有効性の尺度を決める手法が提案されている。しかしながら、ありとあらゆる問題に普遍的に適用できる素性選択の手法は現時点では存在しない。したがって、優れた素性選択の手法を探究することは依然として重要な課題として残されている。

1.2 研究の目的

本研究では、素性選択にニューラルネットワーク（Neural Network; NN）を用いた手法を提案する。NN は機械学習の 1 つであり、脳の神経回路網を模している。ニューロンを表現するノードが幾つも繋がった構造を持ち、出力層と呼ばれるノード群から最終的な回答を出力する。本研究では、ノード間の繋がりの重みに着目する。NN では一般に誤差逆伝搬法により、トレーニングデータからノード間の繋がりの重みを学習する。本研究では、学習後のノード間の重みを素性の評価に利用することで素性選択を行う。また、本手法の有効性を評価するため、テキスト分類をタスクとして、本提案手法による素性選択手法と頻度による単純な素性選択手法を用いたモデルのテキスト分類の正解率を比較する。

1.3 論文の構成

本論文の構成は以下のとおりである。2 章では、関連研究としてテキスト分類、機械学習、素性選択の先行研究について述べる。3 章と 4 章では、提案手法について述べる。3 章では、提案手法の概略の説明とテキスト分類に用いる素性の定義について述べる。4 章では、提案手法の中心となる NN による素性の選択手法について述べる。5 章では、提案手法による評価実験の結果とその考察について述べる。6 章では、本研究から得られた成果のまとめと今後の課題について述べる。

第2章　関連研究

本章では、本研究の関連研究について述べる。

2.1　テキスト分類

テキスト分類とは、自然言語で書かれたデータを内容に応じて事前に決められたカテゴリに自動的に振り分けるタスクである。テキスト分類では、入力するデータの表現形式が分類の精度に大きく影響を与える重要な要素となる。

通常、文書などのデータは $(word_1, word_2, \ldots, word_n)$ のようなベクトル表現で表される。$word_i$ は素性であり、文中に出現する単語あるいはフレーズを表す。この時、$word_i$ には重みが付与され、この重みを決める様々な手法が提案されている。

例として、Lan らによる単語の重みの決定方法 [8] を紹介する。まず Lan らは表 2.1 と表 2.2 に記載した重み付け手法を調査した。表 2.1 は単語の出現頻度を基にした重みであり、対象としている文書内の単語のみから決まる。表 2.2 は文書コレクションにおける単語の出現頻度を基にした重みであり、文書分類の対象としている全文書から決まる。χ^2、ig、gr、OR は教師情報として文書の正しいカテゴリを要する重み付け手法であり、このような教師情報を用いた手法は教師情報を用いない手法より有効と考えられる。Lan らは表 2.1 のような単語の出現頻度に基づく重みと、表 2.2 のような教師情報に基づく重みを組み合わせた $tf.rf$ という手法を提案している。これは、文書内の単語の出現頻度を表す tf と文書コレクションにおける単語の出現頻度を反映した rf との積により重みを求めている。ここではテキスト分類をあるカテゴリに該当する（ポジティブ）もしくは該当しない（ネガティブ）の 2 値に分類することを目標としている。rf は式 (2.1) のように定義される。式 (2.1) において、a は素性 (単語) が出現しかつポジティブに分類される文書数、c は単語が出現しかつネガティブに分類される文書数を表す。この式は、単語が出現する文書集合内でのポジティブとネガティブに分類される文書数の比を表している。なお、rf の算出には a、c を求めるために教師情報、すなわち正しいカテゴリの情報を必要とする。Lan らは提案手法の評価のために、7 つの重み付け手法との比較を行った。比較した重みは $binary$、tf、$tf.idf$、$tf.\chi^2$、rf、$tf.lg$、$tf.logOR$ である。テキスト分類の実験を複数のカテゴリについて行っているため、F 値のマイクロ平均 (micro averaged F) とマクロ平均 (macro averaged F) を評価尺度とした。実験では、文書コレクションとして Reuter 文書を用い、学習アルゴリズムとしてサポートベクターマシンを用いた時に最も良い結果が得られ、F 値のマイクロ平均 F が 0.9272、マクロ平均は 0.90 であった。他の重み付け手

法と比較すると、提案手法は他の重み付け手法よりも高い F 値を示したが、場合によっては $tf.idf$ などの幾つかの手法と F 値があまり変わらないこともあった。

$$rf = log(2 + \frac{a}{max(1, c)}) \tag{2.1}$$

表 2.1: 1 文書のみからの重み付け (Lan 2009)

binary	文書内で出現すれば 1、でなければ 0
tf	文書内の出現頻度
$log(tf)$	$log(1 + tf)$
ITF	$1 - \frac{1}{1+tf}$

表 2.2: 文書集合からの重み付け (Lan 2009)

idf	$log(\frac{N}{n_i})$
idf_{prob}	$log(\frac{N-n_i}{n_i})$
χ^2	χ^2 分布
ig	information gain
gr	gain ratio
OR	Odds Ratio

Altınçay らは単語の出現頻度を用いた新しい文書表現 [1] を提案している。この文書表現は、テキスト分類の精度の向上のため、単語 (素性) にカテゴリ分類に対する有効性の情報を持たせている。カテゴリ分類に対する有効性とは、ここでは、ある素性を用いて文書のカテゴリを分類したとき、その分類の信頼性から算出する。

この研究が対象とするテキスト分類では、テキストをカテゴリ (ポジティブ) とそれ以外 (ネガティブ) の二値へ分類する。このため、素性にポジティブ、ネガティブ、一般性の 3 つのクラスに対する有効性の情報を持たせる。一般性は、ポジティブ、ネガティブに関係なく分類に与える影響を表す。最初に、素性をポジティブまたはネガティブのどちらかに分類する。素性の分類は素性の出現頻度を用いて行う。まず、各素性について、ポジティブまたはネガティブに分類される文書集合における素性の文書内平均出現頻度を求める。その後、ある文書内での素性の出現頻度をポジティブとネガティブの文書内平均出現頻度と比較し、素性をその差が小さい方に分類する。一方、この分類の信頼性を式 (2.2) から求める。$s_{nmc}(t_k)$ は素性 t_k の信頼性を表している。$\hat{f}(t_k)$ は素性 t_k の出現頻度を表している。$\mu_{pos}(t_k)$、$\mu_{neg}(t_k)$ はそれぞれポジティブ・ネガティブに分類された文書集合における t_k の平均文書内出現頻度である。i_{pos}、i_{neg} は、それぞれ $\hat{f}(t_k) - \mu_{pos}(t_k)$、$\hat{f}(t_k) - \mu_{neg}(t_k)$ の逆数である。式 (2.2) は 3 つの場合に分けて信頼性の算出を行っている。1 つ目は素性 t_k が出現していないとき、信頼性を 0 とする。2 つ目は素性 t_k が出現しかつ $\hat{f}(t_k)$ が $\mu_{neg}(t_k)$ よりも $\mu_{pos}(t_k)$ に近い場合で、このときの信頼性を $\frac{i_{pos}}{i_{pos}+i_{neg}}$ とする。3 つ目は素性 t_k が出現しかつ $\hat{f}(t_k)$ が $\mu_{pos}(t_k)$ よりも $\mu_{neg}(t_k)$ に近い場合で、このときの信頼性を $\frac{i_{neg}}{i_{pos}+i_{neg}}$ とする。また、s_{nmc} は素性が出現しないときを除いて、値は 0.5〜1.0 の範囲になる。この値が 1.0 に近いほど、素性の分類の信頼性が高いことを示している。次に、文書ベクトルに持たせる素性の表現について述べる。素性には 3 つの有効性の情報を持たせるため、$\boldsymbol{r_i} = [r_1, r_2, r_3]$ と表される。$\boldsymbol{r_i}$ は素性 i を表している。r_1、r_2、r_3 はそれぞれ、単語 i のポジティブ、ネガティブ、一般性の度合いを示している。この 3 つの重みは素性の分類結果から、式 (2.3) または式 (2.4) により求める。素性がポジティブに分

類されているとき、式 (2.3) で求め、そうでなければ式 (2.4) で求める。それぞれの重みは式 (2.2) で算出した信頼性 $s_{nmc}(t_k)$ を基に与えられる。$w(t_k)$ は出現頻度を用いた重みである。また、素性がポジティブに分類されるときはネガティブの重み、ネガティブに分類されるときはポジティブの重みを 0 にすることで、反対のカテゴリへの影響を抑制している。最終的に文書ベクトルは $[\boldsymbol{r_1}, \boldsymbol{r_2}, \ldots, \boldsymbol{r_K}]$ として表される。Altınçay らは、バイグラム、すなわち隣接する 2 つの単語の組み合わせも素性に使用している。バイグラムでは、$\boldsymbol{r_1^c} = \boldsymbol{r_1} + \boldsymbol{r_2}$, $\boldsymbol{r_2^c} = \boldsymbol{r_2} + \boldsymbol{r_3}$, $\ldots$, $\boldsymbol{r_{K-1}^c} = \boldsymbol{r_{K-1}} + \boldsymbol{r_K}$ によって素性を求める。すなわち、バイグラムの素性を 2 つの単語素性の和で表しており、文書ベクトルは $[\boldsymbol{r_1^c}, \boldsymbol{r_2^c}, \ldots, \boldsymbol{r_{K-1}^c}]$ として表される。

実験では、通常のベクトル、すなわち素性の重みを出現頻度とするベクトルとの比較を行い、提案手法による文書ベクトルは通常のベクトルより高い精度を示した。最後に Altınçay らは、素性に隣接した単語の組み合わせに制限せず同文書に出現した単語の組み合わせを用いてテキスト分類の精度を改善させた Figueiredo らの研究 [4] に触れ、精度の改善のために最適な単語の組み合わせを選択する手法の調査は重要だと述べている。

$$s_{nmc}(t_k) = \begin{cases} 0, & \hat{f}(t_k) = 0 \\ \frac{i_{pos}}{i_{pos}+i_{neg}}, & \hat{f}(t_k) > 0 AND |\hat{f}(t_k) - \mu_{pos}(t_k)| < |\hat{f}(t_k) - \mu_{neg}(t_k)| \\ \frac{i_{neg}}{i_{pos}+i_{neg}}, & \hat{f}(t_k) > 0 AND |\hat{f}(t_k) - \mu_{pos}(t_k)| \geq |\hat{f}(t_k) - \mu_{neg}(t_k)| \end{cases} \tag{2.2}$$

$$\boldsymbol{r_i} = [r_1, r_2, r_3] = [s_{nmc}(t_k) \times w(t_k), 0, (1 - s_{nmc}(t_k)) \times w(t_k)] \tag{2.3}$$

$$\boldsymbol{r_i} = [r_1, r_2, r_3] = [0, s_{nmc}(t_k) \times w(t_k), (1 - s_{nmc}(t_k)) \times w(t_k)] \tag{2.4}$$

テキスト分類のための文書表現として、Cavnar らによって提案されたNグラムを使用した手法 [2] もある。この手法では、単語の両端を示す空白を含む文字単位のNグラムを素性として使用する。例として、単語 text からは、$\_T, TE, EX, XT, T\_$ といったバイグラムの素性が生成される。文書をベクトルとして表現するときは、複数のNグラムを同時に使用する。Cavnar らは $N = 1$〜5 までのNグラムを使用している。素性の重みは tf、すなわち文書内におけるNグラムの出現頻度とする。Cavnar らはベクトルをプロファイルと呼んでいる。未知のテキストのプロファイルをカテゴリのプロファイルと比較して、一番距離の近いカテゴリに分類する。カテゴリのプロファイルは事前にカテゴリの付与された文書から作成する。プロファイル間の距離の測定は out-of-place 法を用いている。この手法では、素性 (Nグラム) を出現頻度により降順に並べたとき、N グラムのプロファイル内での位置が重要な役割を果たす。2 つのプロファイル内でのNグラムの位置の差から文書間の距離を測定している。たとえば、Nグラム TE が 2 つのプロファイル内で、それぞれ左側から 2、4 番目の位置にあるとき、位置の差は 2 となる。位置の差を全Nグラムについて求め、それらを合計した値が 2 つのプロファイルの距離となる。Cavnar らは、様々なコーパスで実験を行い、タイプミスの多い E-mail の文書や短い文の分類に提案手法が有効に働くことを示した。

2.2 機械学習

機械学習は大量のデータ内から規則やパターンなどの情報を自動的に獲得する技術である。この技術はデータの分類や予測などに用いられる。機械学習アルゴリズムとして様々な手法が提案されている。一般に分類や予測の精度を高めるために大量のトレーニングデータを必要とする。機械学習は使用するデータの種類によって、2つの手法に大別される。

1つめは教師あり学習といわれる。この手法では、分類モデルの学習のために正解が付与されたデータを必要とする。正解が付与されたデータを基にモデルの学習を行うため、学習後のモデルによる分類の正答率は高くなりやすい。教師あり学習の代表的な手法にサポートベクターマシン (Support Vector Machine;SVM) と NN がある。

SVM は、データを高次元のベクトルとして表現し、データがある分類クラスに該当するか否かを判断する二値分類器を学習するアルゴリズムである。クラスに該当するデータを正例、該当しないデータを負例とし、正例と負例を識別するモデルを学習する。正例と負例の識別は図 2.1 のような分離平面をみつけることで行う。図 2.1 における丸は正例、三角は負例を表す。SVM は新たなデータに対する分類の正解率を高めるため、分離平面から最も近いデータ（サポートベクター）との距離 (マージン) が最大となるように分離平面を学習することで汎用性を高めている。一方、正例と不例の線形分離が難しいデータの集合に対しては、カーネル関数によって線形分離可能な空間に写像することで二値分類器を学習する。自然言語処理における多くの先行研究において、SVM は他のアルゴリズムに比べてよい成果を挙げている。鈴木らは文書の要約作成に必要な重要文節の抽出処理において、SVM によって文節が重要か否かの判断を行っている [14]。形態素などの文節情報と文の長さなどの文情報を素性として SVM を学習する。また、SVM によって重要と判定された文節と係り受け木の情報から要約を作成している。実験では、事前に決められた要約率を満たすまで文書の先頭から文の選択を行う LEAD 法との比較を行った。その結果、F 値で LEAD 法より高い値を示した。

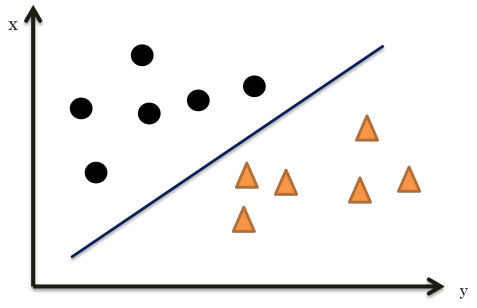

図 2.1: SVM における分離平面

NN は、人間の脳の神経網を模した機械学習の手法である。その基本的な構造は、図 2.2

のようにノードが繋がって構成されたネットワークである。前の層のノードの出力を繋がっている次層のノードに伝播させる。ノードは層に分割され、それぞれ入力層、隠れ層(あるいは中間層)、出力層と呼ばれる。入力層は判定を行うデータの情報を受け取るノードの集合である。隠れ層は入力層と出力層の間にあるノードの集合を指す。出力層は各ノードの出力を NN の結果として返すノードの集合である。隠れ層は複数の層で構成することが可能である。隠れ層のノードの数や各ノード間の繋がりの有無といった NN の構造の差異は、学習した NN による分類の精度に影響を与える。このため、先行研究では様々な構造の NN が提案されている。

Sang らはシステムの異常検知に evolutionary neural network(ENN) と呼ばれる NN を使用した手法 [6] を提案している。ENN は遺伝的アルゴリズムを用いて分類に最適な NN の構造を学習する。NN の構造学習のため、遺伝的アルゴリズムにおける 3 つの構成要素を以下のように決めている。1 つめは Genotype 表現であり、それぞれの NN をどのように表現するかの問題である。ここでは、NN を $N \times N$ の行列で表す。N は入力層、隠れ層、出力層の全ノード数である。行列の要素はノード間の重みである。重みが 0 のとき、ノード間にリンクが存在しないことを表している。2 つめは Genetic 操作であり、交叉と変異を実装している。交叉は 2 つの NN をランダムに選択し、隠れ層のリンク情報の交換を行う。これは 2 つの NN の隠れ層に同じノードが存在するとき、それらから 1 つのノードを選択して、NN 間でそのノードのリンクの重みを入れ替える。変異はリンクの追加または削除を行う操作である。ランダムに 2 つのノードを選択し、それらの間にリンクがある場合はそれを削除し、そうでなければランダムに設定した重みを持つリンクを 2 ノード間に持たせる。3 つめは Fitness Evaluation であり、世代ごとに優良な NN の選択を行う。本研究の NN の選択手法では、NN をランク付けし、それぞれの NN が選択される確率を求める。次世代に残す NN は確率的に選択されるが、ランクが上位の優良な NN ほど高い確率で選択される。NN のランクはトレーニングデータでのシステム異常の検出率から算出する。実験では、NN のリンクの数が 285 個の初期状態から始め、150 まで減らし、トレーニングデータにおけるシステムエラーの検出精度は 90% だった。先行研究の Elman NN との比較では、共にテストデータに対するエラーの検出率 (再現率) は 100% を示したが、誤検出率については提案手法は Elman NN より低かった。

SVM や NN のような教師あり学習の手法の問題の 1 つとして、トレーニングデータ作成のコストが高いことが挙げられる。データに正解を付与する作業に多大だな労力を要するため、多くのトレーニングデータを必要とする機械学習では無視できない問題となる。

機械学習のもう 1 つの手法は教師なし学習である。教師なし学習は、正解の付与されたデータを必要としない手法である。トレーニングデータに正解の情報が含まれていないため、教師あり学習に比べて学習されたモデルによる分類精度は低くなる。一方、トレーニングデータに人手で正解を付与する必要がないため、大量のトレーニングデータを用意しやすいという利点がある。教師なし機械学習は分類問題をはじめとする様々な問題の解決に利用されるが、ここでは一例として教師なしクラスタリングを紹介する。

クラスタリングは 1 つのデータの集合を複数の部分集合に分割する。データ間の類似

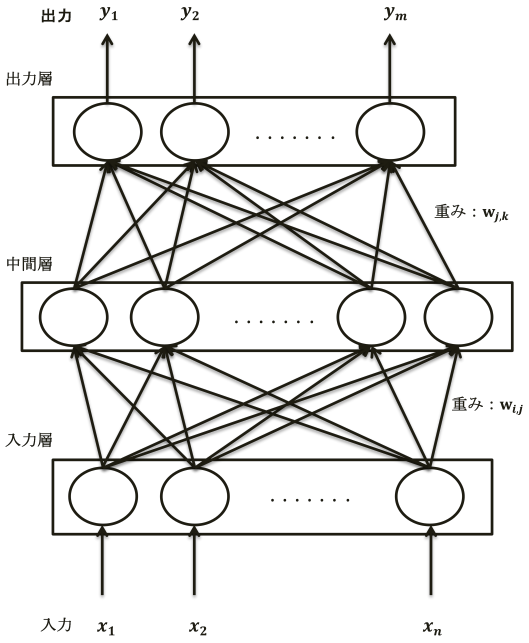

図 2.2: ニューラルネットワークの基本構造

度や距離などの尺度を用いることにより、個々のデータをどの部分集合に割り振るか決める。

クラスタリング手法の 1 つにスペクトラルクラスタリング [12] がある。多くのクラスタリング手法と同様にデータを素性ベクトルで表現し、素性ベクトル間の類似度を基にクラスタリングを行う。ただし、スペクトラルクラスタリングでは、データを表す素性のベクトルをより低い次元の素性ベクトルに変換してからクラスタリングを行う。最初にデータの集合から無向グラフを作成する。グラフのノードはデータを表し、類似しているデータ間には辺を持たせ、データ間の類似度を辺の重みとする。データ間の辺の有無を決める手法には、事前に類似度の閾値を決め、それ以上の類似度をもつデータ間のみ辺を持たせる手法などがある。次に、作成したグラフの辺の類似度から $N \times N$ 個の要素を持つ行列を作成する。N はデータ数であり、i 行 j 列目の要素はデータ i とデータ j の類似度とする。グラフ内で辺を持たないデータ間の要素値は 0 とする。次に、作成した行列の固有値と固有ベクトルを求める。行列の各行をデータの元の素性ベクトルとし、これを固有ベクトルを用いて次元の低い素性ベクトルに変換する。最後に変換後のベクトルを用いて k-Means アルゴリズムなどの従来のクラスタリングによりデータの集合を分割する。スペクトラルクラスタリングは、しばしば k-Means アルゴリズムより高い精度を示す。Hui らはクラスタリングによって、名前の曖昧性を解消する手法を提案している [5]。論文に記載されている名前は、同姓同名の人の存在や略字の使用などにより複数の研究者と合致する。ここで

の名前の曖昧性の解消とは、論文中の名前が指している著者を特定することを指す。Huiらは、名前以外に論文のタイトルや出版雑誌名などの情報を素性ベクトルの要素とし、スペクトラルクラスタリングによって、著者の特定を行った。実験では、著者特定の精度は最高値で 96.8% となり、比較対象とした k-Means アルゴリズムより高い精度を示した。

2.3 素性選択

素性選択は、与えられた素性集合から分類に有効な素性の部分集合をみつける技術である。素性選択のアプローチは大きく分けて 2 つある。

1 つめは feature weight アルゴリズムである。この手法は、それぞれの素性に対し、分類の対象となるカテゴリとの関連度に基づいて重みを算出し、その重みにしたがって素性のランク付けを行う。このランクを基に有効な素性を選択する。

Harun は情報利得 (Information Gain;IG) と主成分分析 (Principal Component Analysis;PCA) 又は遺伝的アルゴリズム (Genetic Algorithm;GA) を組み合わせたハイブリッド型の素性選択手法を提案している [10]。これは 2 つのステップから構成される。まず、IG により重要性の高い素性を絞り込み、次に PCA と GA によって素性の有効性を厳密に評価して素性選択を行う。評価実験では、kNN と C4.5(決定木) の 2 つを用いてテキスト分類を行った。全ての素性を学習に用いる場合と IG のみによる素性選択を行う場合の比較では、IG の方が kNN、C4.5 の両モデルで高い精度を示した。また、IG と GA の組み合わせと IG と PCA の組み合わせは、ともに IG のみより高い精度を示した。これにより、IG と GA 又は PCA を組み合わせる提案手法の有用性が示された。また、IG と GA の組み合わせと IG と PCA の組み合わせを比較したところ、IG と GA の組み合わせのほうが精度が高かった。

2 つめは subset search アルゴリズムである。この手法は、最適または準最適な素性の部分集合を探索する手法である。最初に素性集合が空又は全ての素性を持つ状態から始め、素性集合による学習の結果を評価しながら素性を追加したり削除することで最適な素性の部分集合をみつける。

このアルゴリズムによる研究の 1 つに Setiono らの手法 [9] がある。この手法では、素性選択に N N を使用している。有効な素性集合を探索する際に、素性を取り除いた後の分類精度の低下率を素性の削除の基準としている。提案手法では、まず学習の精度を向上させるために NN の誤差関数に変更を加えている。NN は学習時、正解と出力の誤差を示す誤差関数の値が最小になるように重みの修正を行う。このとき、分類に重要でない素性に繋がるリンクの重みは小さい値に修正される。Setiono らは誤差関数の Cross-entropy 関数にペナルティ項を設けることで誤差の修正率を向上させ、学習の精度を高めている。式 (2.5) はペナルティ項を追加した Cross-entropy 関数である。k、C は、それぞれトレーニングデータの数、出力層の数を表している。t_p^i はデータ i での出力層 p に期待される正解の出力であり、ここでは 0 又は 1 である。S_p^i はデータ i での出力層 p の出力結果である。右項の $P(w)$ は追加されたペナルティ項であり、式 (2.6) により求められる。h、n は、そ

れぞれ隠れ層のノード数、入力層のノード数である。w_l^m は入力層のノード l と隠れ層のノード m の間のリンクの重みである。β、ϵ_1、ϵ_2 は、それぞれペナルティ項の調節を行う変数である。ϵ_1 と ϵ_2 は素性選択時に決められる。式 (2.5) は、正解との誤差の他に、入力層と隠れ層との間の重みの大きさも考慮されている。ペナルティ項を追加することにより、全体的に重みが高いとき、誤差関数の値が高くなるため重みの修正も大きくなり、重要でない素性の重みはより 0 に近い値へと修正される。

$$-(\sum_{i=1}^{k}\sum_{p=1}^{C} t_p^i \log S_p^i + (1-t_p^i)log(1-S_p^i)) + P(w) \tag{2.5}$$

$$P(w) = \epsilon_1(\sum_{m=1}^{h}\sum_{l=1}^{n} \frac{\beta\ (w_l^m)^2}{1+\ \beta\ (w_l^m)^2}) + \epsilon_2(\sum_{m=1}^{h}\sum_{l=1}^{n}(w_l^m)^2) \tag{2.6}$$

Setiono らの素性選択手法は、素性集合から素性を削除していくことで部分集合の探索を行っている。最初に、データをトレーニングデータと検証データに分割し、初期状態として全素性の集合を持つ NN を基準の NN とする。基準の NN をトレーニングデータで学習し、検証データでの分類精度を最大の分類精度として記録する。次に以下の処理を基準の NN に対して行う。

1. 素性を 1 つ取り除いた NN を N 個作成し、それぞれの NN について、トレーニングデータと検証データで分類精度を求める
2. N 個の NN のランク付けを行う。
3. 素性選択を行う
   (a) ランクから NN_k を選択
   (b) NN_k の学習
   (c) 分類精度の減少率の計算
   (d) NN の更新判断
   (e) 基準となる NN の更新
   (f) 素性選択の終了判断

1. において、N は現在の素性の数である。作成した NN は、ある 1 つの素性の入力ノードに繋がったリンクの重みのみ 0 に変更することで、分類時にその素性を無効化している。2. では、トレーニングデータでの分類精度によって NN のランク付けを行う。また NN の分類精度の平均を求める。3. では、まず 2. のランクにおける上位の NN_k を 1 つ選択する ((a) の処理)。NN_k は素性 k を取り除いた NN とする。(b) では NN_k の学習を行う。(c) では事前に記録していた検証データでの最大の分類精度と NN_k の検証データでの分類精

度の差を求めている。(d) では、(c) の分類精度の差が低いとき (NN_k の分類精度が全素性を用いたときと比べてあまり低下しないとき)、素性 k は分類に有効でないと判断して (e) の処理を行い、そうでなければ (f) の処理を行う。(e) の基準となる NN の更新では、最初に式 (2.6) のペナルティ項のパラメータを修正する。$\epsilon_1(a)$、$\epsilon_2(a)$ は素性 a に対応するパラメータとする。素性 a を取り除いた NN のトレーニングデータの分類精度が平均の分類精度より高いとき、$\epsilon_1(a)$、$\epsilon_2(a)$ を増加させ、低いときは減少させる。これを取り除く素性 k 以外の全ての素性で行いパラメータを調整する。すなわち、ランクが高い素性は分類に大きな影響を与えないため、ϵ_1、ϵ_2 を大きく設定する。次に、素性の集合から素性 k を取り除き、また NN_k を基準の NN にする。さらに、NN_k の検証データでの分類精度が最大の分類精度を上回ったときには、最大の分類精度の値を更新する。以上の更新を行った後、1. の処理に戻る。前述のように ϵ_1、ϵ_2 を大きく設定することにより、次のステップではランクが最上位の NN_k の分類精度の減少率を小さくする (素性を有効でないと判断されやすくする) 効果が生じる。(f) では、素性選択の終了の判定を行う。現在の NN_k が最もランクが低い NN のとき、今残っている素性は分類に有効として素性選択を終了し、そうでなければ (a) に戻る。評価実験として、2 つの問題と 4 つのデータを用いて、全素性を用いた場合と、提案手法によって選択された素性集合を用いたモデルを比較したところ、後者の方が分類精度が高いことを確認した。

また、Verikas らも NN を使用した素性選択手法を提案している [11]。Setiono らの手法との違いとしては、誤差関数への追加項と素性の削除に使用するランキングの求めた方が異なる。式 (2.7) は提案手法における誤差関数の定義であり、cross-entropy 関数に第 2 項と第 3 項が追加されている。E_0 は従来の cross-entorpy 関数で、式 (2.8) で求められる。P はトレーニングデータ数で、Q はカテゴリ数である。d_{jp} はトレーニングデータ p での出力層 j に期待される正解の出力であり、$o_{jp}^{(L)}$ はデータ p での出力層 j での出力値である。n_L、n_h は、それぞれ出力層のノード数、隠れ層のノード数である。一方、式 (2.7) における $\acute{f}(net_{kp}^h)$、$\acute{f}(net_{jp}^{(L)})$ は、それぞれ隠れ層のノード k、出力層のノード j の伝達関数の導関数であり、式 (2.7) の第 2 項、第 3 項はそれぞれ中間層のノードの伝達関数の導関数、出力層のノードの伝達関数の導関数の平均である。また、α_1、α_2 はこれらの影響度を調整するためのパラメータである。これにより、隠れ層と出力層の出力が誤差関数に反映され、これを最小化するように NN が学習される。

$$E = \frac{E_0}{n_L} + \alpha_1 \frac{1}{Pn_h} \sum_{p=1}^{P} \sum_{k=1}^{n_h} \acute{f}(net_{kp}^h) + \alpha_2 \frac{1}{Pn_L} \sum_{p=1}^{P} \sum_{j=1}^{n_L} \acute{f}(net_{jp}^{(L)}) \tag{2.7}$$

$$E_0 = -\frac{1}{2P}[\sum_{p=1}^{P} \sum_{j=1}^{n_L} (d_{jp} \log o_{jp}^{(L)} + (1 - d_{jp}) \log(1 - o_{jp}^{(L)}))] \tag{2.8}$$

次に素性選択の手続きについて述べる。最初に、素性選択時の際に素性を削除する順番を決めるための素性のランキングを求める。また、素性を残すか削除するかを判断するために、NN による分類精度の基準値を求める。素性のランキングと分類精度の基準値は L

個の NN から計算する。一般に、NN の学習は、トレーニングデータと重みの初期値に依存するが、複数の NN を用いるのはその影響を抑えるためである。まず、データをトレーニング、検証、テストデータに分割する。トレーニングデータで NN を学習し、検証データで式 (2.7) におけるパラメータ α_1 と α_2 の調整を行い、テストデータで NN の分類精度を求める。次に、素性の中から、その素性を取り除いたときの分類精度が最も低くなる素性を選択し、その素性を素性集合から削除する。この操作を素性が 2 個以上残っている間継続し、素性を削除した順序を記録する。また、(素性を削除する前の) 全素性集合を用いたときの NN のテストデータの分類精度を求める。以上の手続きを L 個の NN に対して行う。素性のランキングは、L 個の NN における素性の削除順序を基に決定する。一方、分類精度の基準値は、L 個の NN の分類精度の平均とする。以上の手続きを経てから素性選択を行う。初期状態として全素性を持つ NN から始め、素性のランキングの順序にしたがって素性を 1 つずつ取り除いていく。素性を取り除いた NN は、トレーニングデータで学習し、検証データで α_1 と α_2 を調整し、テストデータで NN の分類精度を求める。この分類精度と分類精度の基準値を比較し、事前に定めた閾値より低下していなければ、素性の削除を続け、そうでなければ最後に取り除いた素性を追加して (素性の削除をキャンセルして)、探索を終える。実験では、複数の問題設定で既存のいくつかの素性選択手法と比較し、全ての場合において提案手法が他の素性選択より高い精度を示した。また、比較的多くの素性を取り除くことができたと述べている。

前節で述べたように、NN 内のネットワーク構造は分類の精度を左右する重要な要素である。このため、Kabir らは素性選択と同時に最適な NN の構造も学習する手法を提案している [7]。提案手法の処理の流れを図 2.3 に示す。最初に素性を 2 つの集合に分割する。N 個の素性を素性間の相関性によりランク付けを行い、具体的には他の素性との相関係数を平均した値の降順にランク付けを行い、上位 $\frac{N}{2}$ 個、下位 $\frac{N}{2}$ 個に分ける。上位の素性の集合を S, 下位の素性の集合を D と呼ぶ。提案手法は少数の素性から出発し、2 つの素性の集合から素性を追加していくことで素性選択を行う。素性の追加は D 内のランクの高い素性から追加する。D が空になれば、S 内のランクの高い素性から追加する。素性選択の初期状態は最小の NN、すなわち入力層のノードが 2 個、隠れ層のノードが 1 個、出力層のノードがカテゴリ数の NN とする。入力層のノードは、S、D の素性集合からランクの高い素性を 1 つずつ取り出したものである。この NN に対し、素性 (入力層のノード) と隠れ層のノードを追加していく。図 2.3 の「素性の選択」では、S または D から素性を 1 つ選択し、NN に追加する。次に、NN をトレーニングする。NN は誤差逆伝播法で学習するが、重みの反復学習を τ 回 (τ はあらかじめ決められた定数) 繰り返す。トレーニング後、検証データでの NN の出力層のノードの出力値の誤差をエラー率として求める。このエラー率が素性追加前のエラー率より事前に定めた閾値より大きいとき、トレーニングをやめて素性選択を終了する。そうでなければ、次にトレーニングの続行が必要かの判定を行う。この判定にもエラー率を使用する。前回のトレーニングでのエラー率と比較し、今回のエラー率が事前に定めた閾値より大きいときには NN のトレーニングに戻り、そうでなければ素性の分類に対する寄与の高さ (寄与率) を計算する。素性の寄与率は検証デー

タでの正解率とする。正解率が前回計算した正解率より高いとき、その素性は分類に有効と判断し次の素性の追加へ移る。そうでなければ、隠れ層にノードを追加し NN のトレーニングに戻る。隠れ層にノードを追加したにも関わらず、正解率が再び前回より低くなれば、ここで追加した素性は分類に有効でないと判断し、その素性が追加する前に NN を戻し、次の素性を追加する。実験では、他の素性選択の手法と複数のデータ集合で比較を行い、同程度または他よりよい精度を示した。

2.4 本研究の特色

本節では、先行研究と本研究との違いについて論じる。最初に、タスクとなるテキスト分類は二値分類とする。すなわち、文書があるカテゴリに該当するか否かの判定を行う。これは、提案する素性選択手法のエラーの分析がしやすいように、比較的単純な問題を設定したためである。次に、文書の特徴ベクトルの素性と重みについて述べる。素性は単語や単語を組み合わせたものを使用する。単語はテキスト分類において最も利用される素性であるため本研究でも採用する。単語の組み合わせは、Altınçay らによって単語のバイグラムが有効な素性であることが示されているため、より高い精度を得るために使用する。Cavnar らのように文字単位の組み合わせを使用しないのは、単語が持つ意味の情報が無視されるためである。次に、重みは二値で表現する。すなわち、素性（単語や単語の組み合わせ）が文書に出現するときは 1 、それ以外は 0 とする。こちらも、本手法ではＮＮによる素性選択の有効性の分析を主な目的としているため、複雑な重み付け手法は用いず、単純に 1 又は 0 の二値での重み付けを行う。

機械学習については、本論文では教師なし学習ではなく、教師あり学習によるテキスト分類ならびに素性選択を研究の対象とする。機械学習のアルゴリズムとしては、分類に使用する分類器は、2.2 節で述べていたようにテキスト分類でも高い精度を示している SVM を使用する。一方、素性選択には NN を使用する。ＮＮは 2.3 節で述べたように素性選択にも使用されている。しかし、それらは subset search アルゴリズムであり、何度も学習を繰り返すため、テキスト分類のような多くの素性を持つタスクには不向きである。本研究では、feature weight アルゴリズムで素性をランク付けして素性選択を行い、その際に NN を使用する手法を探求する。

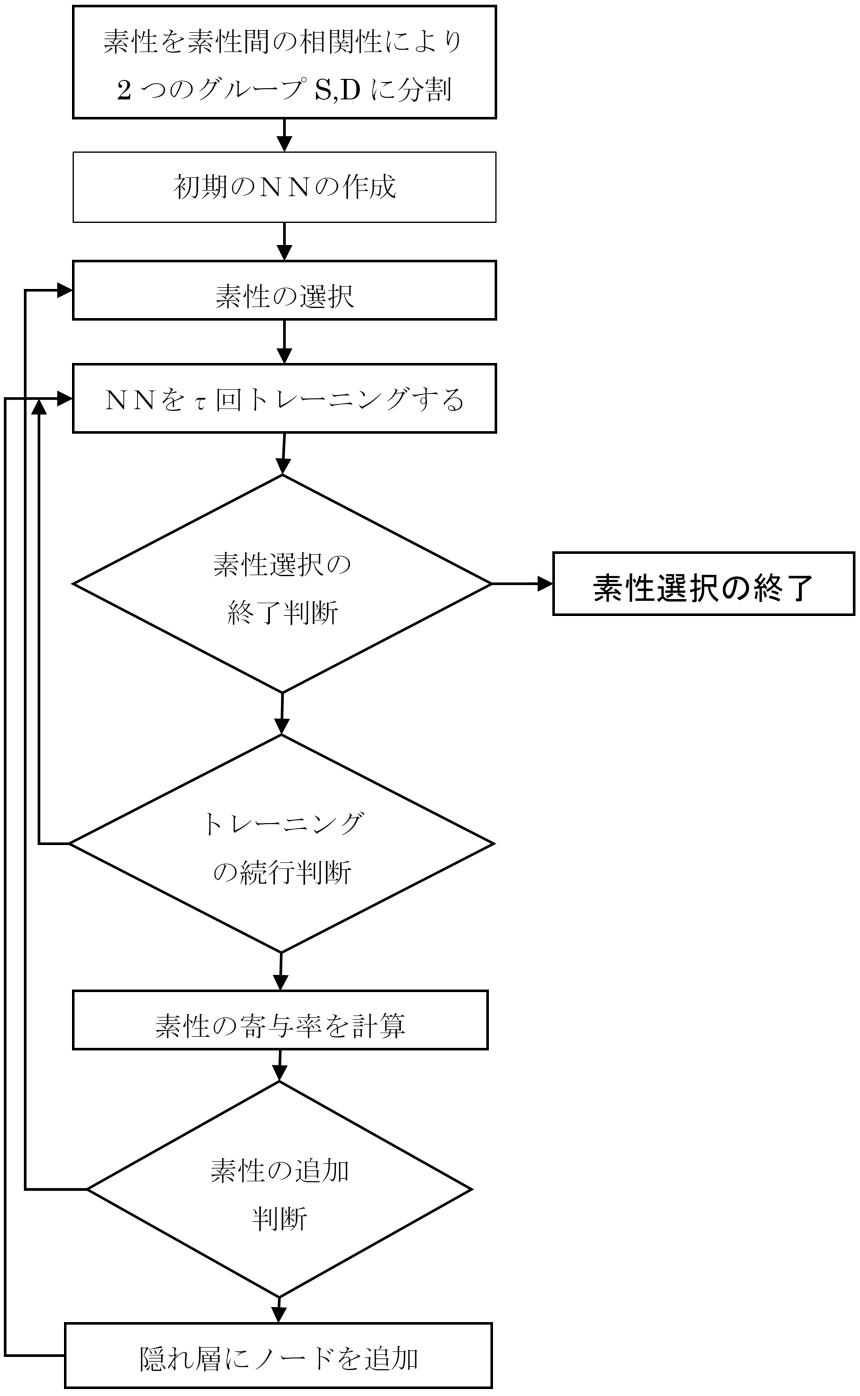

図 2.3: Kabir らによる素性選択手法のフローチャート

第3章　提案手法

本章では提案手法について述べる。まず、本研究において機械学習を適用して問題を解決するタスクとするテキスト分類の概要について説明する。次に、本研究で提案するテキスト分類モデルの概要を述べる。最後に、本研究で使用する学習素性を示す。

3.1　タスク

本論文では、提案する素性選択手法を評価するため、テキスト分類をタスクとする。テキスト分類は自然言語処理分野でも重要な課題の１つである。

テキスト分類は、文書が与えられたとき、その文書のトピックを表すカテゴリを１つまたは複数割り当てるタスクである。この時、カテゴリはあらかじめ人手によって定義されている。例えば、本を分類するためのカテゴリとして、「ＳＦ」、「アクション」、「サスペンス」などをあらかじめ定義する。そして、与えられた文書がＳＦやアクションなどのカテゴリに属するかを自動的に判定する。テキスト分類では教師あり機械学習に基づく手法が主流となっている。機械学習のためのトレーニングデータとして、正しいカテゴリを付与した文書集合を用いる。

本論文では、文書が事前に定義した１つのカテゴリに属するか否かを判定する問題を設定する。このような問題を二値分類といい、入力した文書を２つのクラス (カテゴリに属するか否か) のどちらかに分類する問題である。一方、カテゴリが３個以上のときは多値分類問題と呼ばれる。これは、上記の例のように複数のカテゴリをあらかじめ定義し、その中から本に最も適した「ＳＦ」などのカテゴリを判定するものである。多値分類問題の手法が幾つか知られているが、多くは二値分類以上に多くの計算量を必要とし、評価が難しい。このため、本論文ではカテゴリを１つに絞り、二値分類問題としてのテキスト分類をタスクとして、提案する素性選択手法の評価実験を行う。ここで、分類の対象とするクラスはあらかじめ定義したカテゴリとそれ以外 (Other) の２つである。

二値分類の例を図 3.1 に挙げる。この例は文書を「ＳＦ」カテゴリに属しているか、いないかを判定するテキスト分類である。この時、分類クラスは「ＳＦ」と「ＯＴＨＥＲ」の２つとする。最初に、正解クラスとして「ＳＦ」又は「ＯＴＨＥＲ」を付与した文書集合をトレーニングデータとして用意する。次に、トレーニングデータから「ＳＦ」と「ＯＴＨＥＲ」の特徴を学習し、分類モデル (SF 分類器) を得る。最後に、学習した分類モデルにカテゴリが未知である文書を入力として与えて、「ＳＦ」又は「ＯＴＨＥＲ」のどちらのクラスを持つかを判定させる。

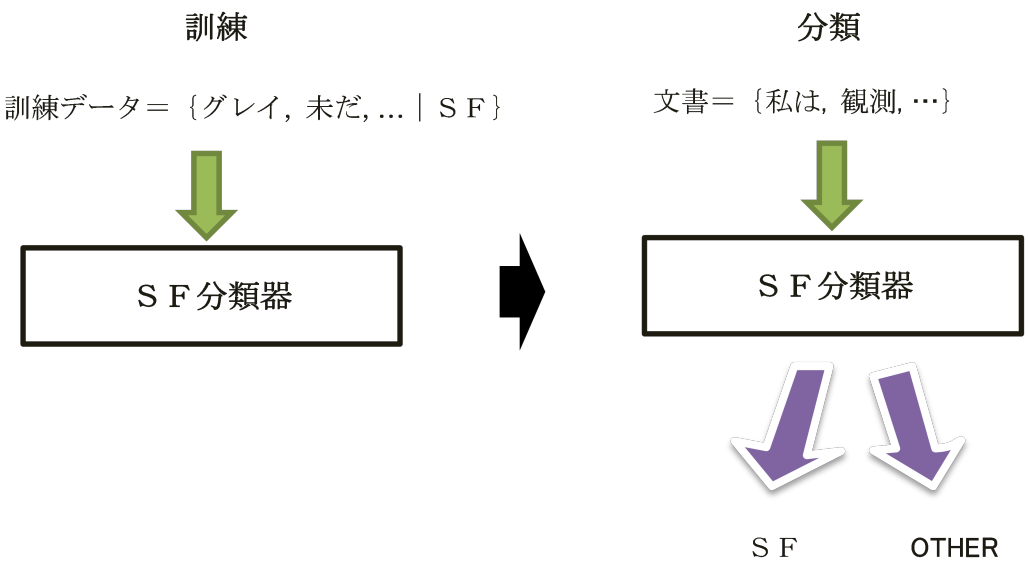

図 3.1: 二値分類のテキスト分類の例

テキスト分類は、一般に文書集合内に出現する単語や単語列を素性とするため、学習素性の数が多いという特徴がある。そのため、テキスト分類では素性選択は重要な処理である。このような理由から、本論文では提案手法の評価実験のタスクとしてテキスト分類を選択した。

3.2 分類モデルの学習

本論文では、ニューラルネットワーク (NN) による素性選択を利用したテキスト分類モデルを提案する。分類モデルを学習するための処理の流れを以下に示す。

1. トレーニングデータから学習素性を抽出
2. 1. で得られたトレーニングデータの学習素性を基にテキスト分類のための NN を学習
3. 2 で学習した NN を基に、学習素性のスコアを計算
4. 3 で得たスコアの上位 T 件の学習素性を選択
5. 4 で選択した学習素性を用いて、テキスト分類のためのサポートベクターマシン (SVM) を学習
6. 未知のデータに対し、学習した SVM によってテキスト分類を行う

学習素性の詳細は次節で述べる。トレーニングデータから学習したNNの重みを基に素性のスコア計算をすることで学習素性のランクを得る。それから、すなわちスコアの高い上位T件の素性から有効な素性集合を構築する。SVMは2.2節で述べたように、最大マージンの学習とカーネルにより、多くの問題で高い汎用性を持つ分類モデルを学習することができる。多くの自然言語処理のタスクで、他の機械学習アルゴリズムと比べて高い精度を得られていることが知られている。このため、本研究ではテキスト分類の学習モデルにSVMを使用している。本論文で提案するテキスト分類モデルの特徴は、素性選択はNNで行い、テキスト分類自体はSVMで行う点にある。

3.3 素性

本節では、本研究で使用する素性について説明する。既に述べたように、素性は機械学習において重要な要素である。本研究では、ユニグラム、バイグラム、共起単語の3種類の素性を用いる。以下、それぞれの定義を述べる。

3.3.1 ユニグラム

Nグラムとは、単語のN個の並びである。Nグラムは多くの自然言語処理のタスクで利用される重要な概念である。ユニグラムはN=1の時のNグラムである。すなわち、本研究で使用するユニグラムの素性とは、文書内に出現する1つの単語を表す。

ユニグラムはテキスト分類において最もよく利用される。例えば、文書がカテゴリ「SF」に属するか否かを判定したい時、science, spaceship, teleportationといった単語は、その文書がSFに関連していることを示唆する。このように、ユニグラムはテキスト分類において有用な素性である。しかし、a, the, onなどの付属語は有効な素性ではない。なぜなら、これらの単語は同じユニグラムでもほとんどの文書に出現し、文書のカテゴリを判定するための手がかりとはならないためである。本実験では図3.3をストップワードとして取り除いている。ストップワードとしては、冠詞以外にも、Iやhimなどの代名詞などもほとんどの文で共通してみられる語として取り除いている。また、文書内に出現する語は複数形などに語形が変形しており、そのままでは同じ語でも別々の語として認識してしまう。このため、出現した語を語幹に戻す処理(ステミング)を行う。本実験は、ストップワードの除去とステミング処理を行った語を素性として抽出する。図3.3に本研究で用いたストップワードのリストを示す。

例えば、図3.2の文書からは、以下のような単語がユニグラムの素性となる。

decline, reveal, bank ...

She declined to reveal the bank's planned response but another bank official said
it has not paid the penalty.
The bank said in a brief statement it believes that its use of the tax-free reserves is legal.

図 3.2: 文書の例

from i want he she her him his her mine my it may would must can could a an the not no and
but or nor yet for besides nevertheless also moreover however still otherwise else therefore
so either neither as that whether if will some who what when why where while since till until
after before soon because unless suppose about away aboard above across against along among
around at behind below beneath beside between beyond by concerinng despite down during per
pro except form in inside into less near of off on onto opposite out outside over round regarding
since than through throughout to toward under underneath up upon via versus abaft absent
apropos astride athwart atop betwixt pace qua sans vice with within without ah aha eh er gee
gosh heh hmm huh oh uh yeah d s t

図 3.3: ストップワードのリスト

3.3.2 バイグラム

バイグラムはN=2 の時のNグラムである。すなわち、バイグラムは隣り合う 2 つの単語の並びを素性とするものである。ユニグラムではストップワードは除いたが、バイグラムではストップワードの除去はしない。これにより、例えば a time と the time は異なる素性として区別される。

例えば、図 3.2 の文書からは以下のようなものがバイグラム素性となる。

(she, decline), (decline, to), (to, reveal), ...

上記の例からも分るように、バイグラムはユニグラムとは異なる性質を持つ。例えば、カテゴリがＳＦか否かを判定するテキスト分類では、time machine は「タイムマシン」という意味を持つので分類に有効な素性となる。しかし、ユニグラムの time, machine では、それぞれ「時間」、「機械」という意味を持ち、文書がＳＦに属するかを示唆しているわけではない。一方、バイグラムには素性の数が組み合わせ的に増大するという問題がある。これはNグラム一般の問題でもあり、Nが大きくなればなるほど素性の数が膨大になる。バイグラムはユニグラムより素性の数が増えることから、特にトレーニングデータの量が少ないときにはテキスト分類の正答率の低下を招く可能性がある。

3.3.3 共起単語

ここでいう共起単語とは、バイグラムと同様に 2 単語の組を素性としたものである。ただし、バイグラムとは異なり、隣り合う 2 つの単語を素性とするのではなく、同じ文書に離れて出現した単語の組を素性とする。また、バイグラムではストップワードも考慮されるが、共起単語の素性では無視される。すなわち、共起単語の素性とは同じ文書に出現する自立語の組である。2 つのユニグラム素性を組み合わせたものともいえる。

共起単語では、同じ文書に出現した単語の組ということでバイグラムとは違う情報を持つ。例えば、文書中に discover と grey がある場合は、grey または discover があるユニグラムとして出現する文書に比べ、ＳＦに属している可能性が高い。バイグラムのように隣接する単語を素性としているわけではないので、必ずしも付属語を含める必要はない。本研究では、ユニグラムと同様に付属語はテキスト分類のための有効な素性ではないと考え、付属語は共起単語の素性として使わない。

例えば、図 3.2 の文書からは次のようなものが共起単語の素性として抽出される。

(declin, reveal), (declin, bank), (declin, planned), . . . , (declin, legal)
(reveal, bank), . . .

第4章　素性選択手法

本章では、提案する素性選択の手法について述べる。最初に提案手法におけるニューラルネットワーク (NN) を詳述する。次に、構築した NN から素性のスコアを計算する方法について述べる。最後に、SVM の学習に用いる素性集合の定義について述べる。素性集合は異なるいくつかの素性選択手法によって作成する。

4.1　ニューラルネットワーク

本論文で素性選択に使用する NN についての詳細を述べる。

4.1.1　構造

本論文における NN の構造を図 4.1 に示す。入力層、中間層、出力層が 1 つずつの 3 層の NN である。入力層のノードはそれぞれが学習素性に対応している。図 4.1 では入力情報となる素性は $feature_1$ として与えられている。文書において、$feature_1$ が出現していれば、それに対応する入力ノードに 1 を与え、そうでなければ入力ノードに 0 を与える。このため、入力層のノード数は対象とする文書集合内の全学習素性数 N となる。出力層のノードはそれぞれがカテゴリに対応している。本研究では、タスクとして二値分類を想定しているため、出力層には対象カテゴリとそれ以外 (Other) の 2 個のノードが存在する。最後に中間層について述べる。中間層のノード数は処理時間や学習の精度に大きく影響を与える。一般に中間層のノード数が多いほど学習に時間を要する。また、中間層のノード数は少ないと学習精度が悪くなるが、ノード数が多すぎても学習後の NN の分類精度は下がってしまう。本研究では中間層のノードの数を $\frac{N+2}{2}$ としている。このノード数は、入力層と出力層の合計の半数である。一方、ノード間のリンクは、図 4.1 に示したように、入力層と中間層、中間層と出力層のノード間に対してのみ張られている。また、リンクは 2 層間の全てのノードと繋がっているため、そのリンク数は $P \times Y$ となる。P は前層のノード数、Y は次層のノード数である。すなわち、入力層と中間層の間には $N \times \frac{N+2}{2}$ 個のリンクが、中間層と出力層の間には $\frac{N+2}{2} \times 2$ 個のリンクが存在する。

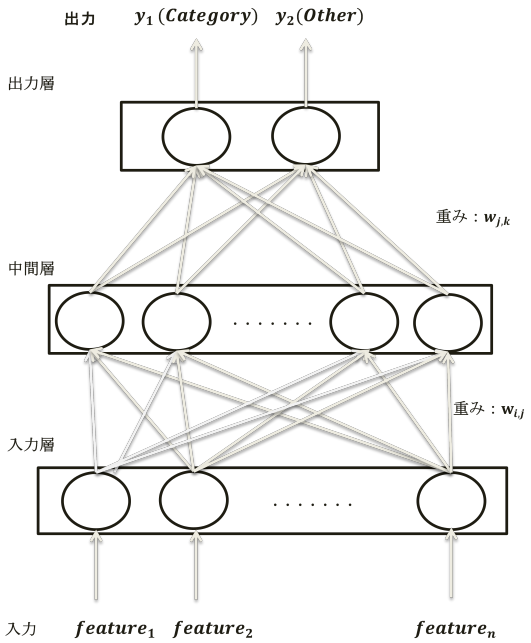

図 4.1: ニューラルネットワークの構造

4.1.2 ニューラルネットワークの学習

NN によるテキスト分類は以下のように行われる。まず、分類対象のテキストから得られる素性に応じて入力層に入力が与えられる。入力層、中間層、出力層という順序で信号が伝えられる。NN では、ノードの出力値を次ノードに渡す時、値を調節してから次ノードの入力値として渡す。この際、値の調節を行う変数がリンクの重みである。式 (4.1) は前層のノード群からの入力値を表す。i は前層の入力ノードの番号、j は当該ノードの番号を表している。M は j とリンクでつながれている前層のノード数、w_{ij} はノード i からノード j への重みを表しており、x_i はノード i の出力値を表している。式 (4.1) の値は式 (4.2) の識別関数に渡され、ノード j の出力値を決定する。ξ はしきい値を表しており、式 (4.2) は入力値 f_j が ξ を越えていたら 1、以下なら 0 を出力する。式 (4.1) と式 (4.2) から分かるように、重みは次ノードの出力値の決定に影響を与えている。最後に出力層のノードの信号を調べ、テキストは 1 を出力するノードに該当するカテゴリに分類される。

NN の学習は、ノードの重み w_{ij} を学習することである。本論文では誤差逆伝播法 [13] により w_{ij} を学習する。誤差逆伝播法は、正解の出力と NN からの出力の誤差が小さくなるように重みを反復的に調整するアルゴリズムである。重みは式 (4.3) で計算される誤差に応じて調整される。式 (4.3) において、N は出力層のノード数で、d_j は出力層のノード i での正解の値で、f_i は出力層のノード i が出力した値である。誤差の影響は中間層と出力層の間のリンクの重みから、入力層と中間層の間のリンクの重みと伝えられる。

$$f_j = \sum_{i=1}^{M} w_{i,j} x_i \tag{4.1}$$

$$y_j(f_j) = \begin{cases} 1, & (f_j > \xi) \\ 0, & otherwise \end{cases} \tag{4.2}$$

$$E = \frac{1}{2} \sum_{j=1}^{N} (d_j - f_j)^2 \tag{4.3}$$

4.2 素性のスコア付け

本節では、本手法で提案する素性のランク付けに使用する素性のスコアについて述べる。ここでは 3 種類のスコアを提案する。それぞれ、入力・中間モデル ($socre_A$)、入力・中間＋中間・出力モデル 1($score_B$)、入力・中間＋中間・出力モデル 2($score_C$) と呼ぶ。

4.2.1 入力・中間モデル

ここでは入力層と中間層間のノードの重みを利用した $score_A$ の定義について述べる。式 (4.4) は $score_A$ の計算式であり、素性 i のスコアを、その素性に対応する入力層のノードとそれに繋がる中間層のノード間のリンクの重みの平均とすることを表している。

$$score_A(i) = \frac{1}{M} \sum_{j=1}^{M} w_{ij} \tag{4.4}$$

M は中間層のノード数であり、w_{ij} は入力層のノード i と中間層のノード j とのリンクの重みである。式 (4.4) は、ノード i が中間層のノードに与える影響の大きさを定量化したものであり、このスコアが大きい素性ほど重要な素性とみなしている。図 4.2 は $feature_i$ のスコアの計算に使用するリンクの重みを表している。すなわち、$feature_i$ に対応するノード i と中間層のノードとのリンクの重みのみによってスコアを求めている。

4.2.2 入力・中間＋中間・出力 1 モデル

ここでは、入力層と中間層間のノードの重みに加え中間層と出力層間のノードの重みも利用した $score_B$ の定義について述べる。式 (4.5) は $score_B$ の計算式であり、$score_A$ と比べて中間層と出力層間のリンクの重みの和が追加されている。

$$score_B(i) = \frac{1}{2M} \sum_{j=1}^{M} w_{ij} \times \sum_{k=1}^{2} w_{jk} \tag{4.5}$$

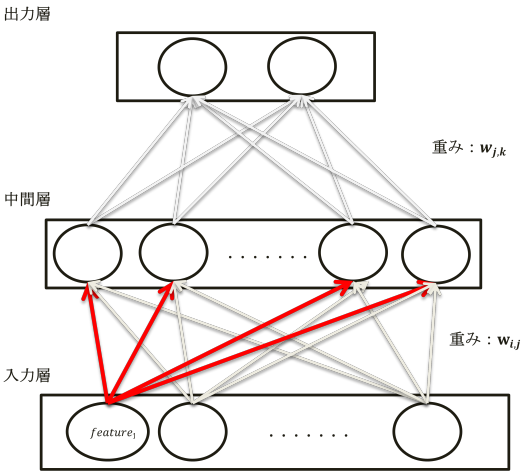

図 4.2: $score_A$ の計算に用いるリンク

w_{jk} は中間層のノード j と出力層のノード k とのリンクの重みを表している。$score_B$ はノード i から中間層を経て出力層のノードへ到達するパス上の重みの積 $(w_{ij} \times\ w_{jk})$ の平均である。全体を $2M$ で割っているのは、上述のパスの数が $2M$ だからである。$score_A$ はノード i とそれに繋がる中間層のノードの間のリンクの重みによって素性 i が出力に与える影響を測っているのに対し、$score_B$ ではノード i とそれに繋がる中間層のノード、さらにその中間層のノードと出力層の間のリンクの重みによって素性 i の影響度を定量化している。$score_B$ が高いほど、すなわち出力層のノードに与える影響が大きいほど、その素性は重要な素性であるとみなす。図 4.3 は $score_B$ による $feature_i$ の計算に使用するリンクを表している。$feature_i$ に対応するノードと中間層のノードとのリンク、および中間層のノードと出力層のノード間の全リンクの重みが $score_B$ の計算に用いられる。

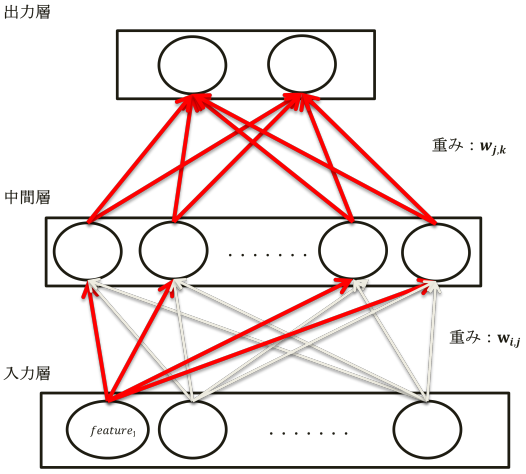

図 4.3: $score_B$ の計算に用いるリンク

4.2.3　入力・中間＋中間・出力２モデル

ここでは、$score_B$ のように入力層と中間層、中間層と出力層の間のリンクの重みを用いるが、後者の利用法が異なるモデル $score_C$ について述べる。式 (4.6) は $score_C$ の計算式である。

$$score_C(i) = \frac{1}{M}\sum_{j=1}^{M} w_{ij} \times max_k w_{jk} \tag{4.6}$$

この式では、中間層から出力層への２つのリンクのうち、重みが高い方のみをスコアの計算に用いている。全体を M で割っているのは考慮している入力層から出力層へのパスの数が M 個だからであり、パス上のリンクの重みの平均値を素性のスコアとしている。これは、２つの出力ノードのうち正しいのは１つのみであることから、不正解と思われる出力ノードをスコアの計算に用いないことで素性の有効性をより正確に測ることを狙っている。図 4.4 は $score_C$ による $feature_i$ のスコアの計算に用いるリンクを示している。中間層から出力層へのリンクは重みが高い方のリンクしか選択されていないことに注意して頂きたい。

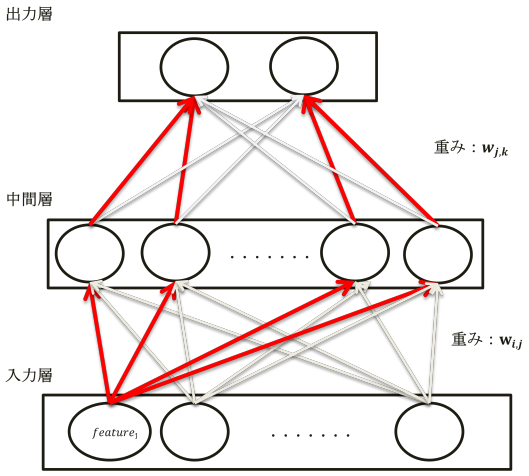

図 4.4: $score_c$ の計算に用いるリンク

4.3　素性集合の定義

本節では、本手法で SVM の学習素性として使用する素性集合の定義について述べる。素性 (単語ユニグラム、単語バイグラム、共起単語) の有効性や NN による素性選択手法の有効性を評価するため、以下の 8 種類の素性集合を作成する。

1. **高頻度のユニグラム**
   この素性集合は、単語ユニグラムの集合である。ここでは、ユニグラムの出現頻度によって素性選択した素性集合を作成する。すなわち、出現頻度が M 以上のユニグラムによって素性集合を作成した。M が小さいほど素性集合のサイズは大きくなり、M が大きければサイズは小さくなる。

2. **高頻度のユニグラム＋バイグラム**
   ユニグラムとバイグラムを合わせた素性集合である。本素性集合では出現頻度が 5 回を超えるもののみを使用している。すなわち、ユニグラムで 6 回以上出現した単語、バイグラムでは 6 回以上隣接して出現した単語の並びを素性としている。

3. **高頻度のユニグラム＋共起単語**
   ユニグラムと共起単語を合わせた素性集合である。本素性集合でも出現頻度が 5 回を超えるもののみを使用している。すなわち、6 回以上同じ文書に出現した 2 つの単語の組を共起単語の素性とする。

4. **高頻度かつ NN で素性選択されたユニグラム**
   出現頻度が 5 回を超えるユニグラムのうち、NN によって得られる素性のスコアの上位 N 個を選択した素性集合である。素性のスコアは 4.2 節で述べた $score_A$、$score_B$、$score_C$ のいずれかで計算している。

5. **高頻度かつ NN で素性選択されたユニグラム＋バイグラム**
   出現頻度が 5 回を超えるユニグラムもしくはバイグラムのうち、NN によって得られる素性のスコアの上位 N 個を選択した素性集合である。素性のスコアは $score_A$、$score_B$、$score_C$ のいずれかで計算している。

6. **高頻度かつ NN で素性選択されたユニグラム＋共起単語**
   出現頻度が 5 回を超えるユニグラムもしくは共起単語のうち、NN によって得られる素性のスコアの上位 N 個を選択した素性集合である。素性のスコアは $score_A$、$score_B$、$score_C$ のいずれかで計算している。

7. **NN で素性選択されたユニグラム**
   NN によって得られる素性のスコアの上位 N 個のユニグラムを選択した素性集合である。素性のスコアは $score_A$、$score_B$、$score_C$ のいずれかで計算する。4. の素性集合とは異なり、出現頻度によって素性の選別を行っていない。すなわち、出現頻度が小さい素性も NN で得られる素性のスコアが高い素性は素性集合に加えれられる。

8. **高頻頻度のユニグラム＋素性選択された共起単語**
   この素性集合は、出現頻度が 5 回を超えるユニグラムと、NN によって得られる素性のスコアが上位の単語の組み合わせの共起単語から構成される。後者の共起単語の素性は以下の手続きで選別する。まず、NN の素性のスコア ($score_A$、$score_B$、$score_C$

のいずれか) の大きい上位 N 個の単語ユニグラムを選別する。次に、得られたユニグラム (単語) を他の単語と組み合わせて共起単語の素性を得る。同一単語を除く全ての単語の組み合わせを作るため、共起単語の素性の数は $N \times$ (全素性数 $- 1$) となる。

第5章　評価実験

本章では、提案手法を評価するために行った実験について述べる。5.1 節では実験に使用したコーパスについて述べる。5.2 節では、評価実験の手続きを説明する。5.3 節では、実験結果と考察について述べる。

5.1　コーパス

本実験では、英語コーパスである Reuters text collection を用いる。このコーパスはロイター社が発信した 21578 個の電子ニュース記事で構成されている。Reuters text collection はテキスト分類の研究に利用する事を前提にしたコーパスであり、本文以外にも様々な情報が付与されている。それぞれの文書には、5 種類のカテゴリが付与されている。5 種類のカテゴリは、それぞれ複数のサブカテゴリによって細分化されている。本研究では、5 種類のカテゴリのうち TOPIC カテゴリに着目する。このカテゴリのサブカテゴリは経済のトピックに関する分類となっている。本実験では、TOPIC カテゴリにおける出現頻度が上位 10 個のサブカテゴリをテキスト分類のカテゴリとして使用する。10 個のサブカテゴリを表 5.1 に示す。本実験におけるタスク設定は、それぞれのカテゴリについて、与えられたテキストがそのカテゴリに該当するかを判定することである。

表 5.1: 対象とする 10 個の TOPIC サブカテゴリ

earn, acq, trade, crude, money-fx, wheat, corn, grain, interest, ship

本実験では、Reuters text collection のうち 8550 個を利用する。これは、出現頻度が 5 回を超える単語を少なくとも 1 個以上含む文書のみを実験に使用したためである。表 5.2 は、10 個のカテゴリについて、それぞれのカテゴリが付与されている文書数を示している。文書数が多いのは earn と acq で、2000 個以上の文書に付与されている。その他のカテゴリの文書数は 200〜500 個程度である。

Reuters text collection では 1 つの文書に複数のカテゴリが付与されていることがあるが、本実験のタスクは二値分類であり、カテゴリの判定を行うときは、対象カテゴリ以外のカテゴリは無視する。すなわち、8850 個の文書のうち、対象カテゴリが付与されている文書を正例、それ以外の文書を負例とする。したがって、正例と負例の割合はカテゴリ

によって異なる。本実験では、データを 4 : 1 に分割し、それぞれトレーニングデータ、テストデータとする。

表 5.2: サブカテゴリ別の文書数

サブカテゴリ	文書数
earn	3776
acq	2210
money-fx	684
grain	574
crude	565
trade	515
interest	424
ship	295
wheat	287
corn	224

5.2 実験手順

本節では、評価実験の手順を述べる。本研究では、素性選択に NN を、テキスト分類に SVM を用いる。これらの機械学習は Weka[1] を用いて行った。プログラムは、Eclipse[2] 上で実装した。

本実験は、図 5.1 で示すように、文書データから素性の抽出、素性の選択、テキスト分類、評価という順序で行う。最初の処理は、文書から素性を抽出する (図 5.1 の 1)。3.3 節で述べたように、本手法では単語を素性とするため、文書から単語を抽出する。このとき、ストップワードの除去と Stemming 処理を行う。Stemming は PaiceHuskStemming[3] の手法を採用した。2 つの処理を行った後、単語をユニグラム素性として抽出する。バイグラムは Stemming 処理のみを行い、2 単語の連続を素性として抽出する。次に、図 5.1 の 2 では、文書をベクトルで表現する。ベクトルの次元は素性、各次元の重みは素性が文書に存在するときに 1、それ以外は 0 とする。図 5.1 の 3 では、素性選択のために NN を学習する。今回の実験環境では全素性を用いた NN はメモリ不足により学習することができなかった。そのため、素性をいくつかの部分集合に分割し、それぞれの素性集合で小さな NN を学習した。1 つの NN の素性数を 125 とし、素性の分割はランダムに行った。本来は全ての素性を用いた NN を学習することで、複数の素性の有効性を同時に評価するべきである。今回の実験では素性を 125 個ずつのグループに分けて NN を学習するため、同

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] https://www.eclipse.org/

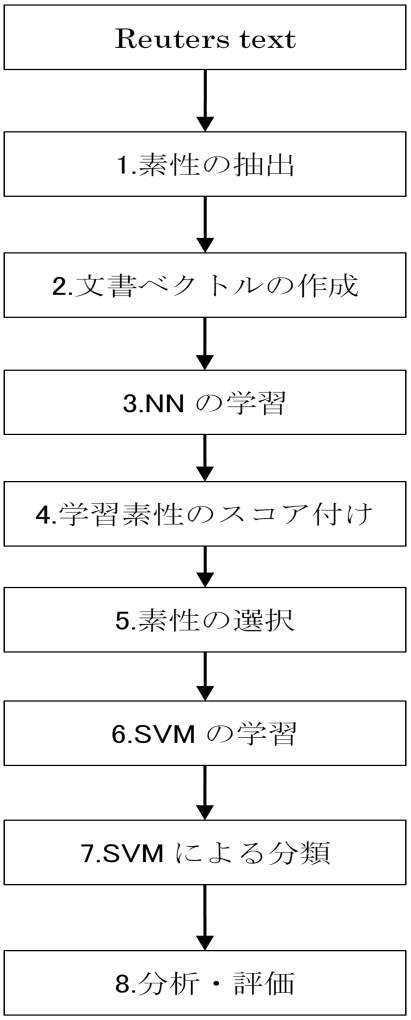
Reuters text
1.素性の抽出
2.文書ベクトルの作成
3.NN の学習
4.学習素性のスコア付け
5.素性の選択
6.SVM の学習
7.SVM による分類
8.分析・評価

図 5.1: 実験手順

じグループ内の 125 個の素性間の有効性は厳密に比較できるが、異なるグループに属する素性間の有効性は厳密に比較できない。実験環境の制約から、素性の有効性を近似的に評価せざるを得なかった。図 5.1 の 4 では、学習素性のスコア付けを行う。学習をした NN のリンクの重みの情報を用いて、4.2 節で述べた計算法からそれぞれの学習素性のスコアを見積もる。学習素性はスコアの値の降順に並びかえる。図 5.1 の 5 では、学習素性の選択を行う。学習素性の選択では、スコアの上位の素性集合を選択する。図 5.1 の 6 では、素性選択後の素性集合を用いてテキスト分類のための SVM を学習する。SVM は Weka で学習する。本実験では、学習時のオプションはデフォルト設定のままとする。図 5.1 の 7 では、学習した SVM を用いてテストデータのテキスト分類を行う。最後に図 5.1 の 8 では、SVM によるテキスト分類の結果を評価し、考察する。

テキスト分類の評価指標として、テストデータでの正答率、精度、再現率、F 値を用いる。正答率は式 (5.1) で定義され、システムの判定結果が正解と一致した割合である。精度、再現率、F 値は、それぞれ式 (5.2)、式 (5.3)、式 (5.4) により定義される。精度は、システムがカテゴリに該当すると判定したとき、どれくらいの割合の文書が実際にカテゴリに該当するかを評価する。再現率は、テストデータ内に存在する正例 (カテゴリに該当する文書) のうち、どれだけシステムによって正しく分類されたかを評価する。F 値は、精度と再現率の調和平均である。一般に精度と再現率はトレードオフの関係にあるため、F 値は両方を総合的に評価するときに用いられる。

$$\text{正答率} = \frac{\text{カテゴリの判定結果と正解が一致した文書数}}{\text{テストデータの文書数}} \tag{5.1}$$

$$\text{精度} = \frac{\text{カテゴリに該当すると正しく判定された文書数}}{\text{システムによりカテゴリに該当すると判定された文書数}} \tag{5.2}$$

$$\text{再現率} = \frac{\text{カテゴリに該当すると正しく判定された文書数}}{\text{テストデータ内におけるカテゴリに該当する文書数}} \tag{5.3}$$

$$F\text{ 値} = \frac{2 \times \text{精度} \times \text{再現率}}{\text{精度} + \text{再現率}} \tag{5.4}$$

本実験では、テキスト分類の結果を上記の 4 つの指標で評価する。

5.3 実験結果と考察

本節では、4.3 節で定義した素性集合による SVM のテキスト分類の結果を報告し、考察を行う。

5.3.1 素性の有効性

最初に、実験で使用した 3 種類の素性の有効性を比較する。比較する素性集合は、高頻度のユニグラム、高頻度のユニグラム＋バイグラム、高頻度のユニグラム＋共起単語であ

る。なお、本項で述べる高頻度のユニグラムは、$M = 5$ の素性集合である。すなわち、6 回以上出現した単語のみを素性としている。

表 5.3 は、高頻度のユニグラム、高頻度のユニグラム＋バイグラム、高頻度のユニグラム＋共起単語の素性集合を用いて学習した SVM のテキスト分類の結果を示している。表 5.3 に示す正答率、精度、再現率、F 値は、10 カテゴリの平均値である。カテゴリごとのテキスト分類の結果は付録 A の表 A.1、A.2、A.3 に示す。正答率では、高頻度のユニグラムは 0.979 であり、高頻度のユニグラム＋バイグラムと高頻度のユニグラム＋共起単語の 0.981 より低い値であった。精度では、高頻度のユニグラム＋共起単語の 0.872 が最も高く、次に高頻度のユニグラムの 0.85 が高かった。高頻度のユニグラム＋バイグラムは 0.813 であり、高頻度のユニグラムより 0.037 低かった。しかし、再現率、F 値では高頻度のユニグラムが最も高い値を示した。再現率では、高頻度のユニグラムは 0.828 であった。これは、高頻度のユニグラム＋バイグラム (0.68)、高頻度のユニグラム＋共起単語 (0.656) を上回り、その差も約 0.15 と大きい。F 値では、高頻度のユニグラムは 0.838 であり、高頻度のユニグラム＋バイグラム (0.734)、高頻度のユニグラム＋共起単語 (0.742) より約 0.1 程度高かった。F 値を基準にすれば、最も良い素性は高頻度のユニグラムであった。次いで、高頻度のユニグラム＋共起単語の F 値が高い。共起単語とバイグラムを比較すると、再現率は高頻度のユニグラム＋バイグラムが高いが、精度は高頻度のユニグラム＋共起単語の方が高く、両者の平均の F 値では結果として共起単語の方が上回った。高頻度のユニグラムが他の 2 つの素性集合と比べて高い値を示した理由としては、素性数の違いが原因と考えられる。高頻度のユニグラムの素性数は 9643 個であるのに対し、高頻度のユニグラム＋バイグラムは 107892 個、高頻度のユニグラム＋共起単語は 129277 個であった。高頻度のユニグラム＋バイグラムと高頻度のユニグラム＋共起単語はともに 10 万以上の素性数であることから、トレーニングデータの量に対して使用する素性の数が多く、過学習を起こしていると考えられる。

表 5.3: 3 種類の素性の有効性の評価

素性集合	素性数	正答率	精度	再現率	F 値
高頻度のユニグラム	9643	0.979	0.85	0.828	0.838
高頻度のユニグラム＋バイグラム	107892	0.981	0.813	0.68	0.734
高頻度のユニグラム＋共起単語	129277	0.981	0.872	0.656	0.742

5.3.2 NN による素性選択手法の比較

本論文では 3 種類の NN による素性選択の手法を提案した。ここではこれらの比較を行う。表 5.4、表 5.5、表 5.6 は、それぞれ 3 種類の素性選択手法 (素性のスコア) によるテキスト分類の結果である。NN による素性選択は多くの計算時間を要するため、表 5.4 と表 5.5 では正例の多い earn、trade、acq の 3 つのカテゴリについてのみ実験を行った。一方、表 5.6 は表 5.1 に示した 10 カテゴリの平均である。各表では 3 もしくは 10 個のカテゴリに対する正答率、精度、再現率、F 値の平均値を示す。また、素性選択の際に選択する素性の数については、小さい素性数と大きい素性数の 2 通りについて実験を行った。なお、カテゴリ毎のテキスト分類の結果は付録 A の表 A.4〜A.9 に示す。

表 5.4 は、高頻度かつ NN により素性選択したユニグラムを素性集合としたときの実験結果である。表中の $score_A$、$score_B$、$score_C$ は素性選択に用いたスコアの式を表す。また、素性数は 2500 個と 5000 個の 2 つのサイズで比較している。素性数が 2500 個のとき、正答率、精度、再現率、F 値の全てで $score_B$ が最も高い値を示した。正答率では、$score_B$ は 0.94 で、二番目に大きい $score_A$ の 0.937 と比べて、0.003 とごく僅かな差しかなかった。精度では、$score_B$ は 0.802 と $score_A$ の 0.791 より 0.011 高かった。再現率では、$score_B$ は 0.783 で、$score_A$ の 0.771 より 0.012 高い。F 値でも、$score_B$ は 0.792 と $score_A$ の 0.781 より 0.011 高い値だった。一方、素性数が 5000 個でも 2500 個と同じく $score_B$ が最も高い値を示した。正答率では、3 種類の手法でほとんど差はなかったが、精度、F 値では比較的大きい差がみられた。精度では、$score_B$ は 0.852 で、$score_A$ の 0.827 より 0.025 高かった。F 値では、$score_B$ は 0.839 で、$score_A$ の 0.824 より 0.015 高かった。再現率では、$score_B$ と $score_A$ の差は 0.006 とわずかであったが、$score_C$ とは 0.017 の差があった。以上のことから、NN によってユニグラムの素性を素性選択したときは、$score_B$ が最も良く、次いで $score_A$、$score_C$ の順であった。また、2 つの素性数を比較したとき、3 種類のスコアのいずれも 5000 個の素性集合の方が高い値を示した。$score_B$ の F 値を比較したとき、素性数が 5000 個のときは 0.839 で、2500 個の場合の 0.792 より 0.047 高かった。

表 5.5 は高頻度かつ NN により素性選択したユニグラム＋バイグラムの素性集合の実験結果である。素性数は 10000 個と 15000 個の 2 つのサイズで比較している。素性数が 10000 個の場合、3 つのスコアの間では大きな差はみられなかった。素性数が 15000 個の場合、$score_B$ が最も高い値を示した。正答率は、3 種類のスコアは同程度の値だった。$score_B$

表 5.4: 素性のスコアの比較 (素性集合が高頻度かつ NN で素性選択したユニグラムのとき)

スコア付け	素性数	正答率	精度	再現率	F 値
$score_A$	2500	0.937	0.791	0.771	0.781
	5000	0.954	0.827	0.82	0.824
$score_B$	2500	0.94	0.802	0.783	0.792
	5000	0.954	0.852	0.826	0.839
$score_C$	2500	0.935	0.789	0.762	0.775
	5000	0.952	0.819	0.809	0.814

の精度は他のスコアと比べて約 0.003 程度高く、再現率は約 0.005 程高かった。F 値では、$score_B$ と $score_C$ の差は 0.005 であった。以上の結果から、$score_B$ が最も高い結果を示した。また、素性数の比較では差はみられなかった。

表 5.5: 素性のスコアの比較 (素性集合が高頻度かつ NN で素性選択したユニグラム＋バイグラムのとき)

スコア付け	素性数	正答率	精度	再現率	F 値
$score_A$	10000	0.963	0.859	0.797	0.825
	15000	0.963	0.859	0.796	0.824
$score_B$	10000	0.963	0.859	0.797	0.825
	15000	0.964	0.862	0.801	0.829
$score_C$	10000	0.963	0.858	0.796	0.824
	15000	0.963	0.858	0.797	0.824

表 5.6 は、高頻度かつ NN により素性選択したユニグラム＋共起単語を素性集合としたときの実験結果である。この表では、earn、trade、acq の 3 つのカテゴリの平均を示した表 5.4、表 5.5 とは異なり、実験で使用した 10 カテゴリ全ての平均を示している。また、素性数は、10000、15000、20000、25000 個の 4 通りの場合を比較した。それぞれの素性数で 3 種類のスコアを比較したとき、3 種類のスコアは正答率、精度、再現率、F 値に大きな差はなかった。素性数を変えた場合でも、テキスト分類の結果に大きな差はみられない。F 値では、最も高い値を示したのは、素性数が 20000 個のときの $score_C$ と素性数が 25000 個のときの $score_B$ で 0.834 であった。

表 5.4、表 5.5、表 5.6 の結果から、3 種類のスコア付け手法での間には明確な差がみられなかった。素性集合がユニグラムのとき、または素性集合がユニグラム＋バイグラムで素性集合が 15000 個のときには、$score_B$ が他のスコアと比べて高い評価値が得られたが、他の集合では大きな差はみられなかった。また、$score_A$ と $score_C$ の間の優劣もはっきり

した傾向がみられない。

表 5.6: 素性のスコアの比較 (素性集合が高頻度かつ NN で素性選択したユニグラム＋共起単語のとき)

スコア付け	素性数	正答率	精度	再現率	F 値
$score_A$	10000	0.978	0.852	0.817	0.833
	15000	0.978	0.852	0.816	0.833
	20000	0.978	0.851	0.815	0.832
	25000	0.978	0.851	0.814	0.832
$score_B$	10000	0.978	0.852	0.816	0.833
	15000	0.978	0.852	0.815	0.832
	20000	0.979	0.853	0.814	0.833
	25000	0.979	0.854	0.816	0.834
$score_C$	10000	0.978	0.852	0.816	0.833
	15000	0.978	0.853	0.814	0.833
	20000	0.978	0.853	0.817	0.834
	25000	0.979	0.853	0.816	0.833

5.3.3 提案手法とベースラインの比較

ここでは、本論文で提案する素性選択手法と、高頻度の素性を選択する単純なベースライン手法と比較する。実験の結果を表 5.7、表 5.8、表 5.9 に示す。この表は、ベースラインならびに 3 種類の素性のスコアのそれぞれについて、正答率、精度、再現率、F 値を示しいている。なお、5.3.2 項の実験と同様に、表 5.7 と表 5.8 では earn、trade、acq の 3 つのカテゴリ、表 5.9 は 10 カテゴリに対するテキスト分類の評価値の平均を示した。

表 5.7 は、高頻度のユニグラム (ベースライン) と高頻度かつ NN で素性選択したユニグラムの実験結果である。後者については、表 5.4 において 5000 個のときが 2500 個のときに比べて結果が良かったことから、表 5.7 では素性数が 5000 個のときの結果を再掲した。これらの実験結果の比較から、高頻度のユニグラムが最も高い評価値を示した。提案手法で最も結果が良かったのは $score_B$ のときだが、ベースラインと比べて正答率、精度、再現率、F 値が、それぞれ 0.016、0.038、0.052、0.045 低い。

表 5.7: 提案手法とベースラインの比較 (ユニグラムのとき)

素性集合	スコア付け	正答率	精度	再現率	F 値
高頻度のユニグラム	—	0.97	0.89	0.878	0.884
高頻度かつ NN で素性選択したユニグラム	$score_A$	0.954	0.827	0.82	0.824
	$score_B$	0.954	0.852	0.826	0.839
	$score_C$	0.952	0.819	0.809	0.814

表 5.8 は、高頻度のユニグラム＋バイグラム (ベースライン) と、高頻度かつ NN で素性選択したユニグラム＋バイグラムの実験結果である。後者については、表 5.5 において素性数が 15000 個のときが 10000 個のときと比べて結果が良かったことから、表 5.8 では素性数が 15000 個のときの結果を再掲した。この表の結果から、高頻度のユニグラム＋バイグラムが最も高い評価値が得られた。提案手法は、正答率ではいずれもベースラインと同程度の値を示したが、精度、再現率、F 値では比較的低い値だった。特に再現率の差が大きかった。提案手法の中で最も高い $score_B$ の再現率は 0.801 で、高頻度のユニグラム＋バイグラムより 0.066 低かった。F 値では、いずれもベースラインと比べ約 0.05 低い。

表 5.8: 提案手法とベースラインの比較 (ユニグラム＋バイグラムのとき)

素性集合	スコア付け	正答率	精度	再現率	F 値
高頻度のユニグラム＋バイグラム	—	0.97	0.885	0.867	0.876
高頻度かつ NN で素性選択したユニグラム＋バイグラム	$score_A$	0.963	0.859	0.796	0.824
	$score_B$	0.964	0.862	0.801	0.829
	$score_C$	0.963	0.858	0.797	0.824

表 5.9 は、高頻度のユニグラム＋共起単語 (ベースライン) と、高頻度かつ NN で素性選択したユニグラム＋共起単語の実験結果である。後者については、表 5.6 において素性数

が20000個のときの結果が良かったことから、表5.9では素性数が20000個のときの結果を再掲した。この表はearn、trade、acqの3つのカテゴリの平均を示した表5.7や表5.8とは異なり、実験で使用した10カテゴリの平均を示している。また、参考のため、高頻度のユニグラムを素性集合としたときの結果を表5.9の2行目に示す。これは10カテゴリの平均であり、表5.7の結果と異なることに注意していただきたい。提案手法はベースラインと比べて、再現率とF値で向上していることがわかる。提案手法で最も良かったのは$score_C$であるが、$score_C$は再現率、F値ではベースラインをそれぞれ0.161、0.092上回った。3つのうち最も低い$score_A$でも、再現率とF値でベースラインの結果を上回っている。一方、正答率では、$score_C$はベースラインより劣るが、その差は0.003とごく僅かだった。これに対して、精度では、$score_C$はベースラインと比べて0.019低くなっており、やや大きい差がみられる。4つの評価指標で最も重要なのはF値であり、これを基準にすればNNによる素性選択手法は頻度に基づく素性選択手法に比べて優れていると考えられる。

表5.7、表5.8、表5.9の結果から、ユニグラム、バイグラムでは提案手法はベースラインより下回ったが、ユニグラム＋共起単語を素性にしたときはベースラインを上回った。ただし、表5.9より、ユニグラム＋共起単語の素性でF値が最も高いのは、提案手法の$score_C$を素性選択したときで、その値は0.834であった。一方、ユニグラムを素性とし、頻度によって素性選択したときのF値は同じく表5.9より0.838であった。その差は0.004とごく僅かではあるが、提案手法の方が下回っている。

表 5.9: 提案手法とベースラインの比較 (ユニグラム＋共起単語のとき)

素性集合	スコア付け	正答率	精度	再現率	F 値
高頻度のユニグラム	—	0.979	0.85	0.828	0.838
高頻度のユニグラム＋共起単語	—	0.981	0.872	0.656	0.742
高頻度かつ NN で素性選択した ユニグラム＋共起単語	$score_A$	0.978	0.851	0.815	0.832
	$score_B$	0.979	0.853	0.814	0.833
	$score_C$	0.978	0.853	0.817	0.834

5.3.4 素性集合「高頻度のユニグラム＋ NN により素性選択された共起単語」の評価

前項では、NNによる素性選択手法は、ユニグラム＋共起単語を素性選択としたとき、頻度に基づく素性選択手法よりも優位であることが確認されたが、F値の向上はごく僅かであった。F値をさらに向上させるため、4.3項で述べた「高頻度のユニグラム＋ NN により素性選択された共起単語」という素性集合を考案した。ここで、この素性集合の作成方法を再度述べる。まず、高頻度のユニグラムを素性集合に加える。次に、NNによって素性のスコア付けを行い、その上位N_t個のユニグラムを求める。さらに、上位N_t個のユ

ニグラムと他のユニグラムを組み合わせて共起単語の素性を作成し、これも素性集合に加える。

表 5.10 は、高頻度のユニグラムと、高頻度のユニグラムかつ NN により素性選択された共起単語を素性集合としたときの実験結果である。この実験では、素性のスコアとして $score_A$、$score_B$、$score_C$ の 3 種類を用い、また、共起単語を作成するユニグラムの個数 N_t は 25、50、100、125、150 の場合を試した。表 5.10 の結果は 10 カテゴリの平均である。カテゴリごとのテキスト分類の結果は付録 A の表 A.10〜A.14 に示す。高頻度のユニグラム＋ NN により素性選択された共起単語と高頻度のユニグラムを比較したとき、再現率と F 値で大きな差がみられた。高頻度のユニグラム＋ NN により素性選択された共起単語で結果が最も良かったのは N_t が 25 個かつ $score_B$ のときだが、高頻度のユニグラムと比べて再現率と F 値がそれぞれ、0.079、0.047 低かった。一方、正答率も、高頻度のユニグラム＋ NN により素性選択された共起単語の素性集合の方が低いが、その差は 0.005 程度と僅かである。精度については、高頻度のユニグラム＋ NN により素性選択された共起単語の精度は 0.84 以上であり、幾つかのスコアと N_t の組み合わせでは高頻度のユニグラムを上回った。以上の結果から、本研究で提案する高頻度のユニグラム＋ NN により素性選択された共起単語の素性集合の有効性は確認できなかった。

表 5.10: 高頻度のユニグラム＋素性選択された共起単語によるテキスト分類の結果

素性集合	N_t	スコア付け	正答率	精度	再現率	F 値
高頻度のユニグラム	—	—	0.979	0.85	0.828	0.838
高頻度のユニグラム＋素性選択された共起単語	25	$score_A$	0.974	0.863	0.682	0.749
		$score_B$	0.975	0.842	0.749	0.791
		$score_C$	0.974	0.852	0.716	0.773
	50	$score_A$	0.974	0.839	0.718	0.769
		$score_B$	0.975	0.841	0.717	0.768
		$score_C$	0.974	0.845	0.709	0.765
	100	$score_A$	0.974	0.84	0.716	0.769
		$score_B$	0.974	0.842	0.722	0.774
		$score_C$	0.974	0.855	0.716	0.774
	125	$score_A$	0.974	0.841	0.719	0.771
		$score_B$	0.975	0.856	0.721	0.779
		$score_C$	0.975	0.855	0.713	0.772
	150	$score_A$	0.974	0.847	0.723	0.777
		$score_B$	0.975	0.849	0.738	0.784
		$score_C$	0.974	0.844	0.721	0.773

5.3.5　出現頻度による分割学習した提案手法の評価

本項では、前項と異なるアプローチで NN による素性選択手法の F 値の向上を図った。5.2 項で述べた様に、本実験では、計算機のリソースの問題から、素性を複数の部分集合に分割し、部分集合毎に NN を学習する不完全なものである。部分集合への分割はランダムに行っているが、部分集合によって有効な素性やそうでない素性が偏って集まる可能性がある。有効な素性だけを集めた部分集合を用いて NN を学習した際、真に有効な素性のスコアが相対的に低く見積もられる危険性がある。同様に、有効でない素性だけ集めた部分集合では、真に有効でない素性のスコアが相対的に高く見積もられることがありうる。この問題を解決するために、素性をいくつかの部分集合に分割する際、素性の出現頻度を考慮し、高頻度の素性と低頻度の素性を一様に配置する方法を試す。これは、高頻度の素性は有効な素性、低頻度な素性は有効でない素性である可能性が高いと予想されることから、有効な素性や有効でない素性だけを集めた部分集合を作らないようにするための工夫である。ここで、出現頻度による部分集合の分割手法を述べる。最初に、素性を出現頻度の降順に並べる。次に、先頭から素性を 1 つずつ取り出し、各部分集合に 1 つずつ素性を振り分ける。これを全ての素性に対して行うことで部分集合を作成する。ここでは、素性を部分集合へ分割する手法を区別するため、ランダムでの分割を「ランダム分割」、出現頻度を考慮した分割を「出現頻度による分割」と呼ぶ。

表 5.11 は、2 つの分割手法の実験結果である。素性集合は高頻度かつ NN で素性選択したものを用いた。この表では、5.3.2 項と 5.3.3 項の実験と同様、earn、trade、acq の 3 つのカテゴリに対するテキスト分類の評価値の平均を示した。カテゴリ毎のテキスト分類の結果は付録 A の表 A.15 に示す。表 5.11 から、出現頻度による分割はランダム分割と比べ、$score_B$ で向上がみられた。素性数が 2500 個と 5000 個の両方で、出現頻度による分割はランダム分割より正答率、再現率、F 値が高かった。F 値では、出現頻度による分割で最も良かったのは素性数が 5000 個のときの $score_B$ の 0.841 で、同じ条件のランダム分割の 0.839 より 0.002 高かった。同じ素性数のランダム分割の $score_B$ と比べたとき、再現率が比較的高く上昇しており、F 値も上回っている。ただ、$score_A$ と $score_C$ では、出現頻度による分割はランダム分割より劣る。$score_C$ で素性数が 2500 個のときの F 値は、出現頻度による分割は 0.75 とランダム分割の 0.775 より 0.025 低い。また、素性数を比較したとき、5.3.2 項のユニグラムの実験結果と同じく、3 種類のスコアのいずれも素性数が 5000 個の素性集合の方が高い値を示した。表 5.11 の中で最も F 値が高いのは、出現頻度による分割、$score_B$、素性数が 5000 個のときで、0.841 であった。以上から、出現頻度による分割はランダム分割よりも優れた手法であると言える。ただし、上述の最高の F 値でも、高頻度のユニグラムのベースラインの F 値 (0.884、表 5.7) よりも低い。

5.3.6　提案手法の妥当性の検証

本項では、NN によって算出する素性のスコアの妥当性、すなわちスコアの値がどの程度素性の有効性を的確に表しているのかについて検証する。

表 5.11: 分割手法の比較 (高頻度かつ NN で素性選択されたユニグラムのとき)

スコア付け	素性数	分割手法	正答率	精度	再現率	F 値
$score_A$	2500	ランダム分割	0.937	0.791	0.771	0.781
		出現頻度による分割	0.932	0.741	0.764	0.752
	5000	ランダム分割	0.954	0.827	0.82	0.824
		出現頻度による分割	0.938	0.755	0.785	0.77
$score_B$	2500	ランダム分割	0.94	0.802	0.783	0.792
		出現頻度による分割	0.955	0.827	0.832	0.829
	5000	ランダム分割	0.954	0.852	0.826	0.839
		出現頻度による分割	0.957	0.844	0.837	0.841
$score_C$	2500	ランダム分割	0.935	0.789	0.762	0.775
		出現頻度による分割	0.933	0.744	0.755	0.75
	5000	ランダム分割	0.952	0.819	0.809	0.814
		出現頻度による分割	0.939	0.752	0.789	0.77

まず、ユニグラムを素性集合とし、頻度による素性選択を実施した場合において、素性の数を変化させたときのテキスト分類の結果を調べた。出現頻度が M 以下のユニグラムを削除して素性集合を得るが、ここでは M の値を変化させ、様々な素性数の素性集合を用いたときのテキスト分類の評価指標を比較する。図 5.2、図 5.3、図 5.4、図 5.5 は、10 カテゴリのそれぞれについて、素性数が変化したときの正答率、精度、再現率、F 値を表したグラフである。図 5.2 の正答率のグラフでは、10 カテゴリのいずれにおいても、正答率は素性数が変わっても大きな変化はない。一方、図 5.3 の精度のグラフでは、多くのカテゴリに対して、素性数を減らしたとき、精度は向上したり低下したりする。例えば、カテゴリが interest のとき、素性数が 11000 個から 20000 個の間で精度が増加と減少を繰り返しているのがわかる。図 5.4 の再現率と図 5.5 の F 値のグラフでも、acq と earn を除いて、素性数を減らしていったときの再現率もしくは F 値は向上したり低下したりする。もし、頻度による素性選択が妥当なら、すなわち高頻度の素性ほどテキスト分類に有効な素性といえるなら、素性数を減らすにつれて F 値などの評価指標は単調減少する（左下さがりのグラフになる）はずである。図 5.2〜5.5 の実験結果は、頻度による素性選択が必ずしも適切ではないことを示唆する。

次に、同様にユニグラムを素性集合とし、提案手法となる NN による素性選択を実施した場合において、素性の数を変化させたときのテキスト分類の結果を調べた。ここでは、NN によって算出される $score_A$、$score_B$、$score_C$ のそれぞれについて、そのスコアの上位の素性を選択する。素性数を 3000 から 3000 刻みで 27000 まで変化させたときのテキスト分類の評価指標の変化を調べた。また、素性数が 28795、すなわち素性選択をしない場合の結果も調べた。図 5.6、図 5.7、図 5.8 は、$score_A$、$score_B$、$score_C$ のそれぞ

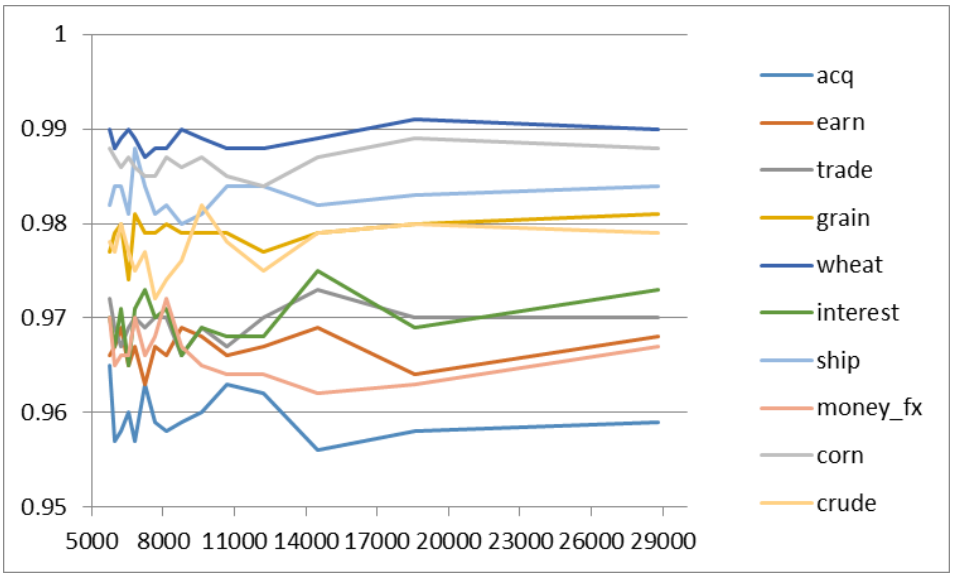

図 5.2: 出現頻度で選択したユニグラムを素性集合としたときの正答率

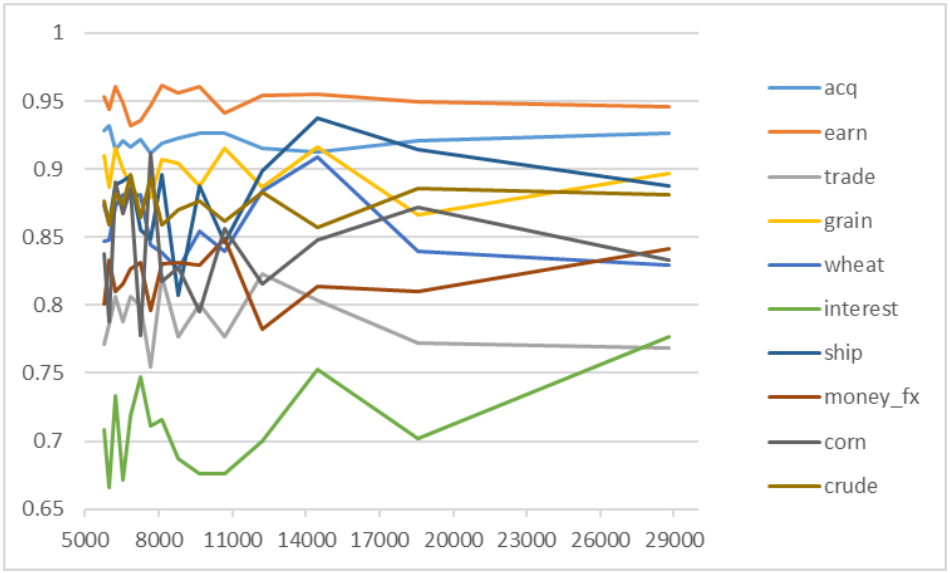

図 5.3: 出現頻度で選択したユニグラムを素性集合としたときの精度

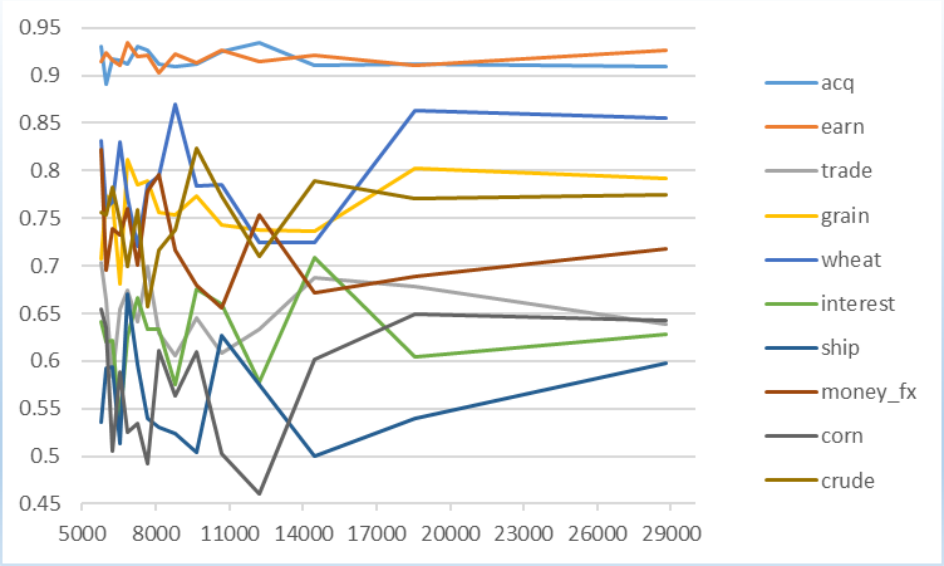

図 5.4: 出現頻度で選択したユニグラムを素性集合としたときの再現率

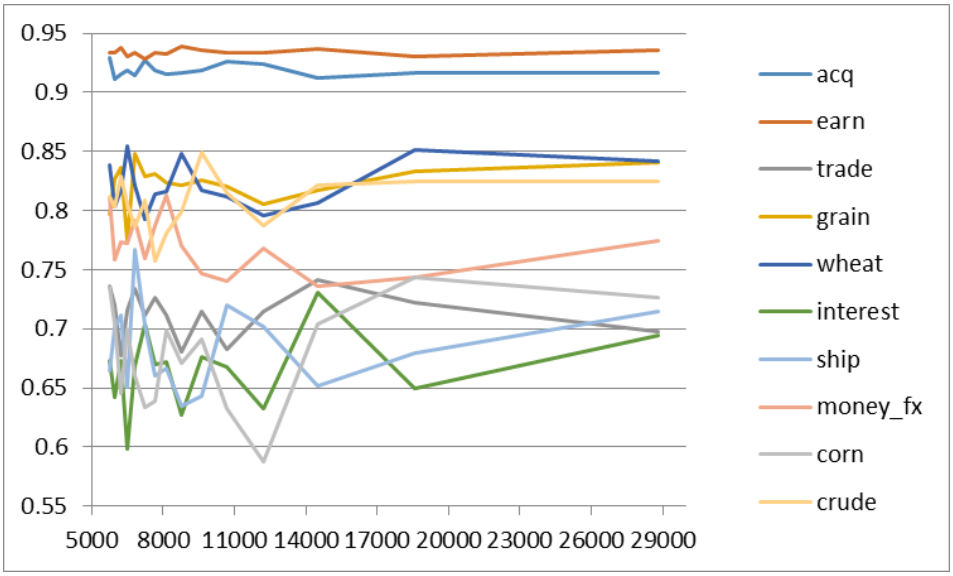

図 5.5: 出現頻度で選択したユニグラムを素性集合としたときの F 値

れについて、素性数を変化させたときの正答率の変化を示している。多くのカテゴリについて、素性数を減らしていくと正答率も単調に減少する傾向がみられる。特に大きな変化がみられるのは acq のカテゴリである。素性数が 3000 個のとき、正答率は $score_A$、$score_B$、$score_C$ でそれぞれ 0.859、0.884、0.858 と低いが、素性数が 9000 個のときには 0.9 以上の値に大きく改善される。図 5.9、図 5.10、図 5.11 は、$score_A$、$score_B$、$score_C$ のそれぞれについての精度の変動を示しいてる。全体的には素性数の減少に伴い精度も減少する。ただし、例外も幾つかみられる。例えば、$score_C$ のグラフ (図 5.11) におけるカテゴリ wheat について、素性数を 9000、6000、3000 と減らしたとき、精度は 0.506、0.519、0.583 と増加している。図 5.12、図 5.13、図 5.14 は、$score_A$、$score_B$、$score_C$ のそれぞれについての再現率の変動を示している。カテゴリ wheat や corn で幾つかの例外がみられるものの、全体的に素性数の減少に伴い再現率は単調減少する。図 5.15、図 5.16、図 5.17 は、$score_A$、$score_B$、$score_C$ のそれぞれについての F 値の変動を示している。やはり幾つかの例外はあるが、素性数の減少に伴い F 値は単調減少するといえる。以上の結果から、$score_A$、$score_B$、$score_C$ の値が大きい順に素性を追加していくと、テキスト分類の結果が改善する傾向がみられる。したがって、提案手法による素性選択のための素性のスコアは、素性の有効性を表す指標として妥当であるといえる。なお、付録 A の表 A.16〜A.25 に、図 5.6〜5.17 でプロットされた点に対応する評価指標の値を示した。

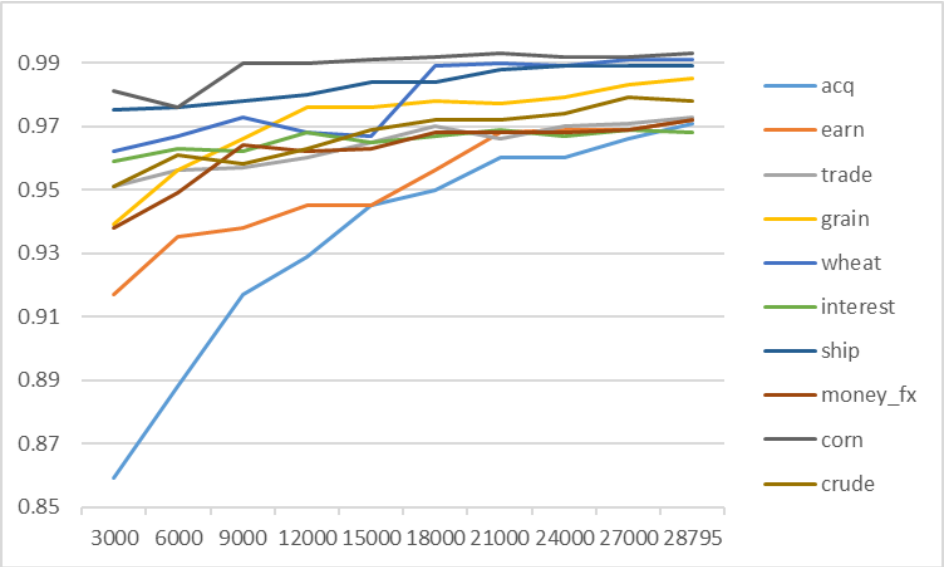

図 5.6: 提案手法 $score_A$ で選択したユニグラムの素性集合の正答率

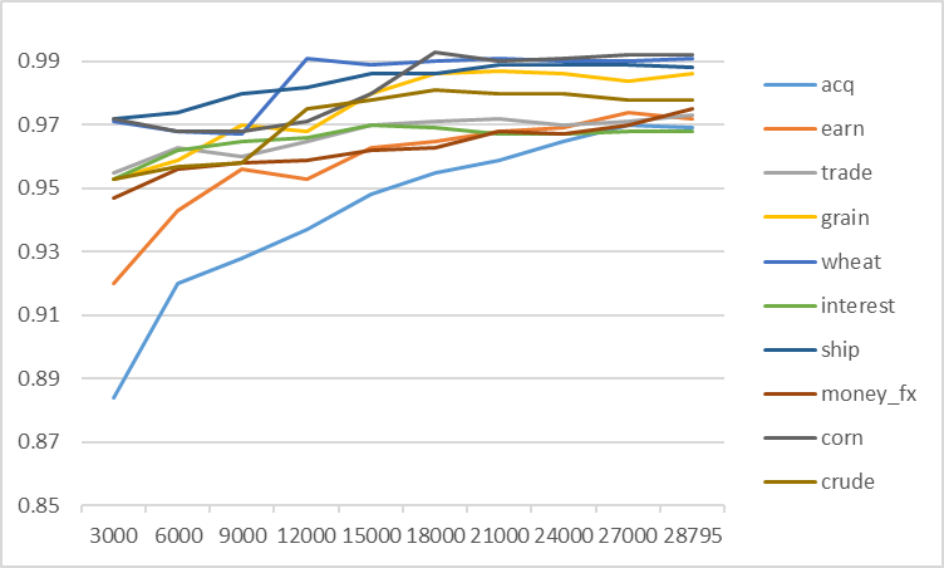

図 5.7: 提案手法 $score_B$ で選択したユニグラムの素性集合の正答率

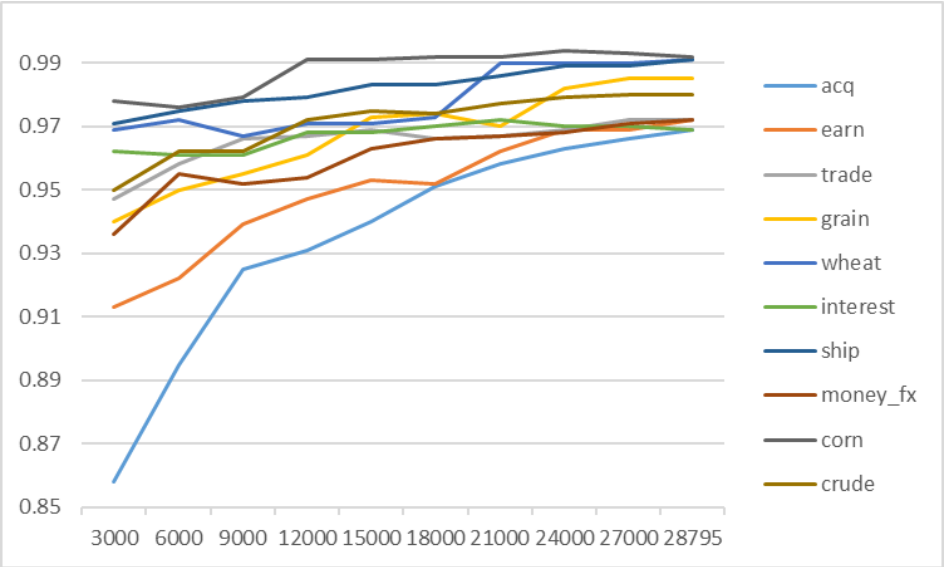

図 5.8: 提案手法 $score_C$ で選択したユニグラムの素性集合の正答率

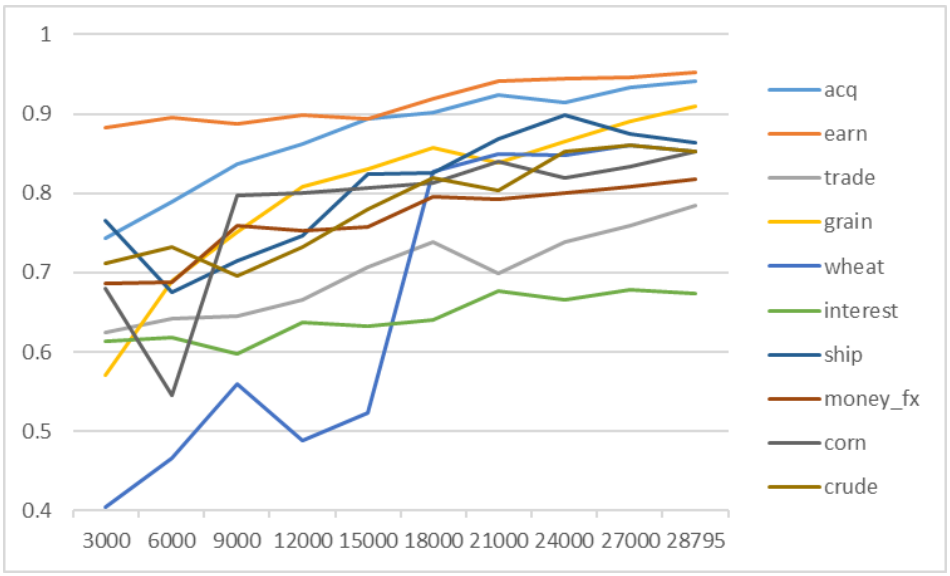

図 5.9: 提案手法 $score_A$ で選択したユニグラムの素性集合の精度

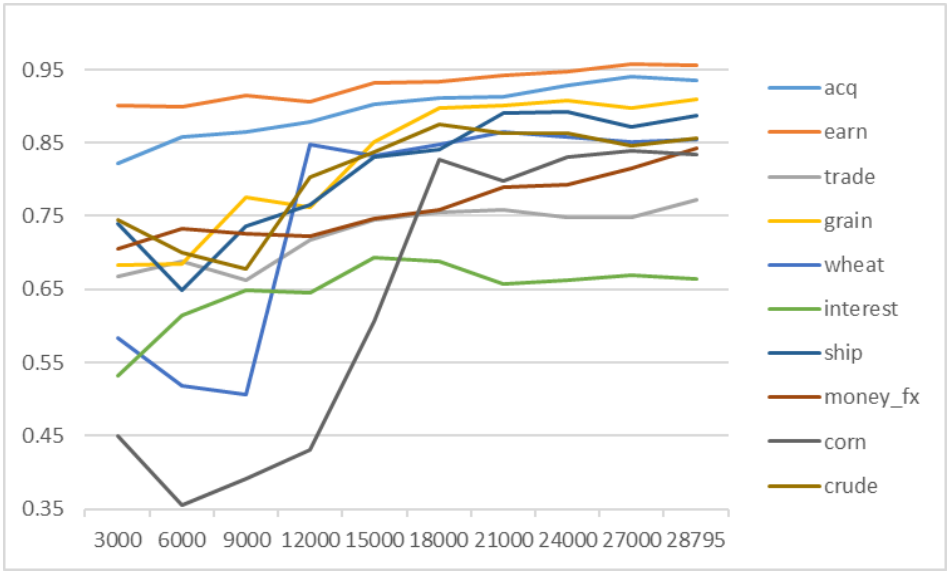

図 5.10: 提案手法 $score_B$ で選択したユニグラムの素性集合の精度

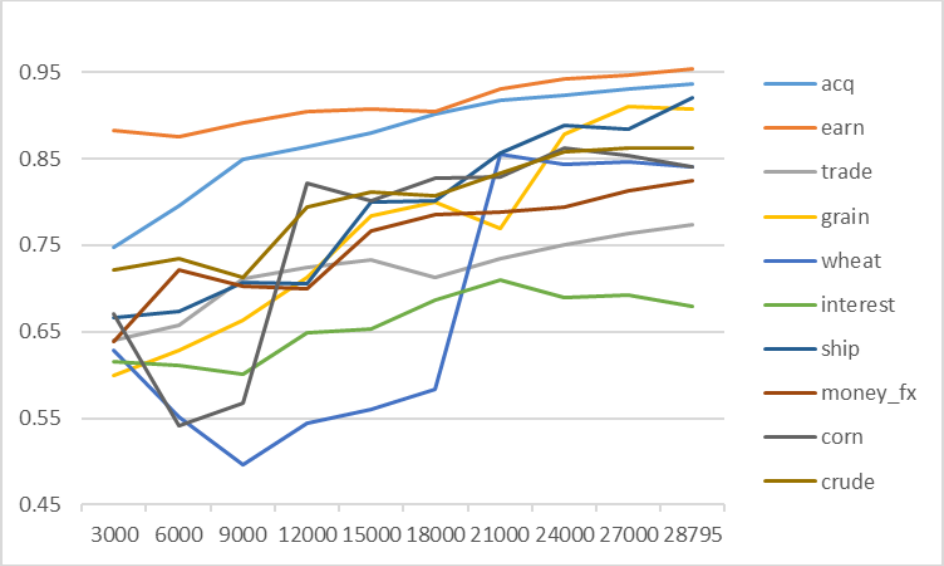

図 5.11: 提案手法 $score_C$ で選択したユニグラムの素性集合の精度

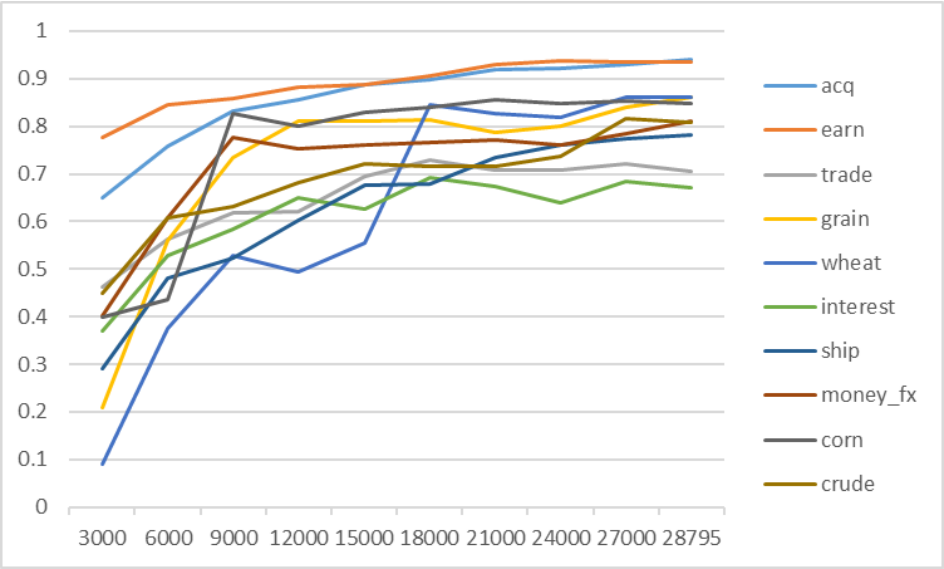

図 5.12: 提案手法 $score_A$ で選択したユニグラムの素性集合の再現率

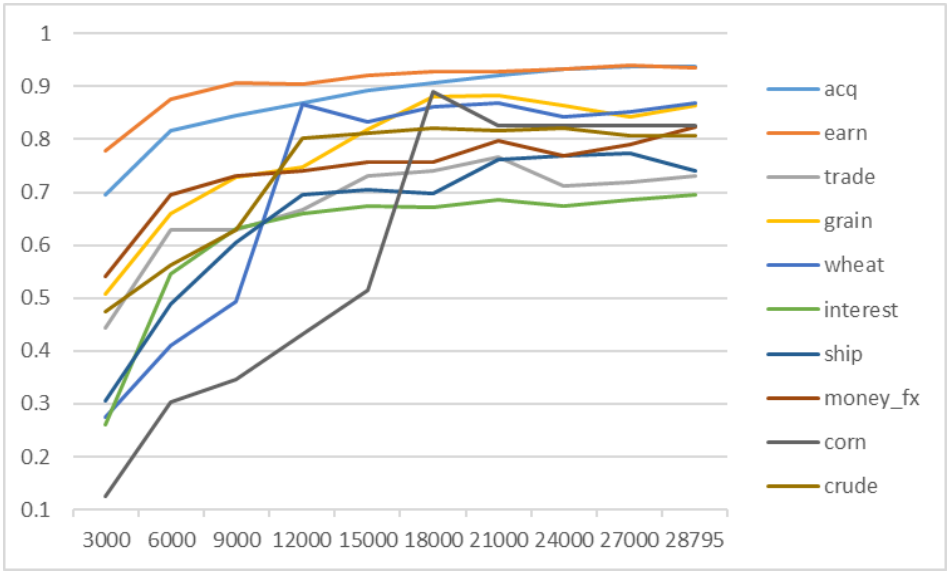

図 5.13: 提案手法 $score_B$ で選択したユニグラムの素性集合の再現率

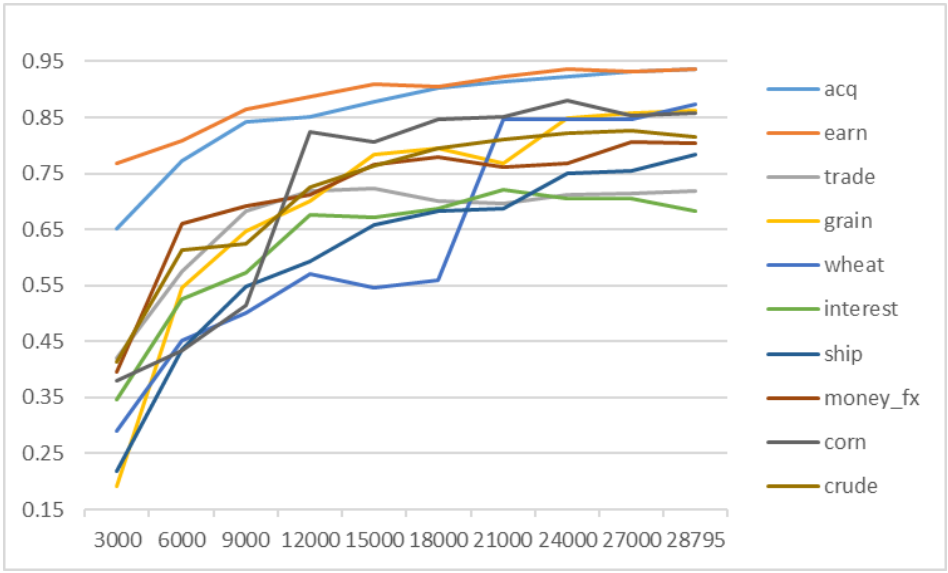

図 5.14: 提案手法 $score_C$ で選択したユニグラムの素性集合の再現率

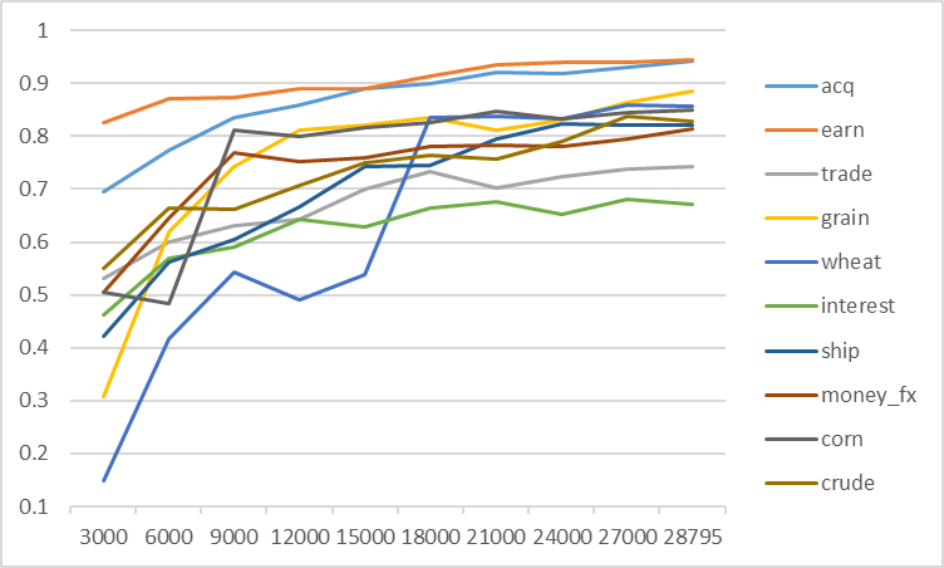

図 5.15: 提案手法 $score_A$ で選択したユニグラムの素性集合の F 値

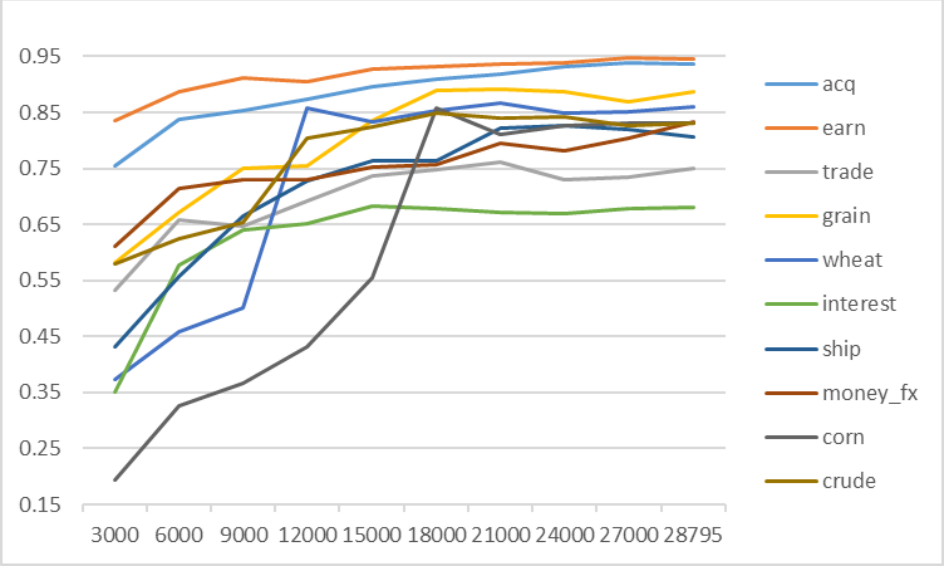

図 5.16: 提案手法 $score_B$ で選択したユニグラムの素性集合の F 値

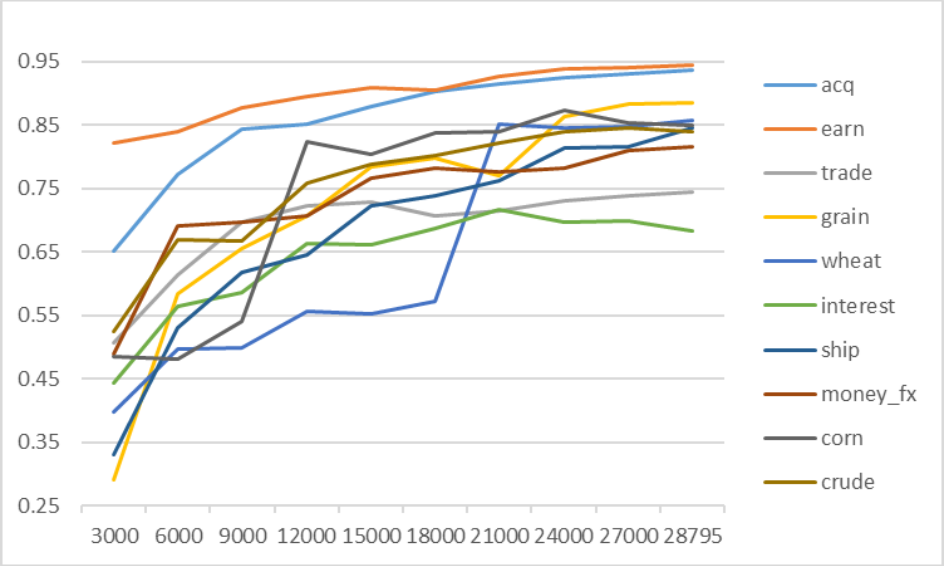

図 5.17: 提案手法 $score_C$ で選択したユニグラムの素性集合の F 値

第6章　結論

6.1　まとめ

本論文では、ニューラルネットワーク (NN) を用いた素性選択手法を提案した。NN 内のノード間のリンクの重みから素性のスコアを計算し、スコアが上位の素性を選択する。素性に対応する入力層のノードと隠れ層のノード間のリンクの重みに基づくものや、入力層から出力層へのパスのリンクの重みに基づくものなど、3 種類のスコアを提案した。素性選択後、サポートベクターマシン (SVM) によって分類モデルを学習した。本論文で対象としたタスクはテキスト分類である。テキスト分類のための学習素性として、単語ユニグラム、単語バイグラムのほかに、文書内に同時に出現する単語の組み合わせである共起単語を用いた。

実験の結果、$score_C$ によって素性選択したユニグラム＋共起単語の素性集合を用いて学習した SVM の F 値は、ユニグラムの素性を単純に頻度によって素性選択したベースラインと同程度であることが確認された。一方、3 種類の素性のスコア付けを比較したところ、明確な差はみられなかった。最後に、NN によって見積もられた素性のスコアの高い順に素性を増やすと、F 値も単調に増加する傾向がみられたことから、提案手法のスコアが素性の有効性を図る指標として妥当であることが確認された。

6.2　今後の課題

今後の課題の 1 つとして、NN の学習時間の問題がある。本研究では、計算機リソースの不足から全素性集合を用いた NN の学習ができなかったために、素性をいくつかの部分集合に分割し、それぞれの部分集合を用いた NN を学習し、部分集合毎に素性の有効性を測っている。このため、素性の有効性を厳密に評価できていない。また、提案手法による素性選択が有効に働く条件を調査することが必要である。実験では、高頻度かつ NN で素性選択されたユニグラム＋共起単語の組み合わせは、単語ユニグラムと比較して、F 値の平均は同程度のものやごく僅かながら向上したものがあった。また、カテゴリ別にみると、高頻度のユニグラムより F 値が向上した組み合わせがあった。このため、どのようなカテゴリあるいは条件のときに NN によって選択された共起単語の素性が F 値の向上に貢献するかを明らかにする必要がある。今回の研究では SVM によって分類モデルを学習したが、ナイーブベイズモデルや決定木などの他の機械学習アルゴリズムと NN による素性選択手法の組み合わせも評価する必要がある。

謝辞

終始熱心なご指導を頂いた主指導教員である白井清昭准教授に感謝の意を表します。島津明教授には、日頃よりご助言を頂きました。ここに感謝いたします。自然言語処理講座の皆様には、研究生活において多くの面で助けて頂きました。この場をお借りしてお礼申し上げます。

参考文献

[1] Hakan Altınçay. Feature extraction using single variable classifiers for binary text classification. In *Recent Trends in Applied Artificial Intelligence*, pp. 332–340. Springer, 2013.

[2] William B Cavnar, John M Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, Vol. 48113, No. 2, pp. 161–175, 1994.

[3] D Paice Chris. Another stemmer. In *ACM SIGIR Forum*, Vol. 24, pp. 56–61, 1990.

[4] Fábio Figueiredo, Leonardo Rocha, Thierson Couto, Thiago Salles, Marcos André Gonçalves, and Wagner Meira Jr. Word co-occurrence features for text classification. *Information Systems*, Vol. 36, No. 5, pp. 843–858, 2011.

[5] Hui Han, Hongyuan Zha, and C Lee Giles. Name disambiguation in author citations using a k-way spectral clustering method. In *Digital Libraries, 2005. JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on*, pp. 334–343. IEEE, 2005.

[6] Sang-Jun Han and Sung-Bae Cho. Evolutionary neural networks for anomaly detection based on the behavior of a program. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, Vol. 36, No. 3, pp. 559–570, 2005.

[7] Monirul Kabir, Monirul Islam, et al. A new wrapper feature selection approach using neural network. *Neurocomputing*, Vol. 73, No. 16, pp. 3273–3283, 2010.

[8] Man Lan, Chew Lim Tan, Jian Su, and Yue Lu. Supervised and traditional term weighting methods for automatic text categorization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, Vol. 31, No. 4, pp. 721–735, 2009.

[9] Rudy Setiono and Huan Liu. Neural-network feature selector. *Neural Networks, IEEE Transactions on*, Vol. 8, No. 3, pp. 654–662, 1997.

[10] Harun Uğuz. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems*, Vol. 24, No. 7, pp. 1024–1032, 2011.

[11] Antanas Verikas and Marija Bacauskiene. Feature selection with neural networks. *Pattern Recognition Letters*, Vol. 23, No. 11, pp. 1323–1335, 2002.

[12] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, Vol. 17, No. 4, pp. 395–416, 2007.

[13] 馬場則夫, 小島史男, 小澤誠一. ニューラルネットの基礎と応用. 共立出版株式会社, 1994.

[14] 鈴木大介, 内海彰. Support vector machine を用いた文書の重要文節抽出-要約文生成に向けて-. 人工知能学会論文誌, Vol. 21, No. 4, pp. 330–339, 2006.

付 録A　カテゴリ毎のテキスト分類結果

5章では、3個もしくは10個のカテゴリの平均を示して考察を行った。付録では、参考のため、各カテゴリ毎のテキスト分類の結果を掲載する。

表 A.1: 高頻度のユニグラムの実験結果

カテゴリ	正答率	精度	再現率	F値
acq	0.968	0.935	0.936	0.935
corn	0.99	0.827	0.818	0.822
crude	0.983	0.885	0.843	0.864
earn	0.972	0.946	0.944	0.945
grain	0.984	0.891	0.868	0.88
interest	0.97	0.702	0.731	0.716
money-fx	0.971	0.811	0.816	0.813
ship	0.987	0.865	0.735	0.795
trade	0.972	0.788	0.755	0.771
wheat	0.99	0.849	0.83	0.84

表 A.2: 高頻度のユニグラム＋バイグラムの実験結果

カテゴリ	正答率	精度	再現率	F 値
acq	0.965	0.929	0.935	0.932
corn	0.993	0.832	0.469	0.6
crude	0.986	0.805	0.712	0.756
earn	0.973	0.952	0.941	0.947
grain	0.986	0.824	0.688	0.75
interest	0.983	0.641	0.565	0.601
money-fx	0.98	0.746	0.686	0.715
ship	0.98	0.819	0.444	0.576
trade	0.971	0.774	0.725	0.749
wheat	0.992	0.809	0.638	0.713

表 A.3: 高頻度のユニグラム＋共起単語の実験結果

カテゴリ	正答率	精度	再現率	F 値
acq	0.96	0.944	0.919	0.932
corn	0.993	0.84	0.497	0.627
crude	0.983	0.939	0.689	0.795
earn	0.969	0.97	0.929	0.949
grain	0.986	0.808	0.69	0.744
interest	0.983	0.892	0.514	0.652
money-fx	0.976	0.755	0.548	0.635
ship	0.99	0.824	0.479	0.605
trade	0.981	0.93	0.649	0.764
wheat	0.992	0.819	0.642	0.72

表 A.4: 高頻度かつ NN で素性選択されたユニグラムの実験結果

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
2500	$score_A$	acq	0.928	0.853	0.859	0.856
		earn	0.922	0.854	0.832	0.843
		trade	0.96	0.666	0.622	0.643
	$score_B$	acq	0.934	0.873	0.864	0.869
		earn	0.926	0.865	0.848	0.856
		trade	0.959	0.667	0.637	0.651
	$score_C$	acq	0.928	0.855	0.851	0.853
		earn	0.92	0.867	0.809	0.837
		trade	0.957	0.646	0.626	0.636
5000	$score_A$	acq	0.946	0.893	0.889	0.891
		earn	0.956	0.926	0.9	0.913
		trade	0.961	0.661	0.672	0.667
	$score_B$	acq	0.946	0.892	0.89	0.891
		earn	0.946	0.901	0.882	0.891
		trade	0.97	0.763	0.707	0.734
	$score_C$	acq	0.944	0.889	0.886	0.888
		earn	0.953	0.923	0.894	0.908
		trade	0.96	0.646	0.646	0.646

表 A.5: 高頻度かつ NN で素性選択されたユニグラム＋バイグラムの実験結果

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
2500	$score_A$	acq	0.962	0.914	0.931	0.923
		earn	0.966	0.964	0.9	0.931
		trade	0.962	0.698	0.559	0.621
	$score_B$	acq	0.961	0.914	0.93	0.922
		earn	0.966	0.965	0.902	0.932
		trade	0.962	0.698	0.559	0.621
	$score_C$	acq	0.962	0.914	0.931	0.923
		earn	0.966	0.964	0.9	0.931
		trade	0.962	0.697	0.557	0.619
5000	$score_A$	acq	0.962	0.915	0.932	0.923
		earn	0.966	0.965	0.901	0.932
		trade	0.961	0.696	0.554	0.617
	$score_B$	acq	0.962	0.914	0.932	0.923
		earn	0.967	0.965	0.902	0.933
		trade	0.963	0.708	0.569	0.631
	$score_C$	acq	0.962	0.914	0.931	0.923
		earn	0.966	0.965	0.9	0.931
		trade	0.961	0.694	0.559	0.619

表 A.6: 高頻度かつ NN で素性選択されたユニグラム＋共起単語の実験結果 (素性数が 10000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
10000	$score_A$	acq	0.967	0.935	0.931	0.933
		corn	0.992	0.853	0.819	0.836
		crude	0.978	0.847	0.814	0.83
		earn	0.971	0.952	0.936	0.944
		grain	0.985	0.91	0.858	0.883
		interest	0.97	0.691	0.701	0.696
		money-fx	0.971	0.823	0.788	0.805
		ship	0.99	0.91	0.77	0.834
		trade	0.969	0.749	0.683	0.714
		wheat	0.991	0.853	0.865	0.859
	$score_B$	acq	0.967	0.934	0.931	0.933
		corn	0.992	0.854	0.825	0.839
		crude	0.978	0.847	0.814	0.83
		earn	0.971	0.95	0.937	0.944
		grain	0.985	0.91	0.858	0.883
		interest	0.97	0.693	0.701	0.697
		money-fx	0.971	0.824	0.783	0.803
		ship	0.99	0.91	0.77	0.834
		trade	0.969	0.748	0.68	0.712
		wheat	0.991	0.853	0.865	0.859
	$score_C$	acq	0.967	0.935	0.931	0.933
		corn	0.992	0.848	0.819	0.833
		crude	0.978	0.847	0.814	0.83
		earn	0.971	0.952	0.936	0.944
		grain	0.985	0.91	0.854	0.881
		interest	0.97	0.69	0.698	0.694
		money-fx	0.971	0.822	0.788	0.805
		ship	0.99	0.914	0.77	0.836
		trade	0.969	0.749	0.683	0.714
		wheat	0.991	0.853	0.865	0.859

表 A.7: 高頻度かつ NN で素性選択されたユニグラム＋共起単語の実験結果 (素性数が 15000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
15000	$score_A$	acq	0.967	0.934	0.933	0.933
		corn	0.992	0.85	0.831	0.84
		crude	0.978	0.847	0.814	0.83
		earn	0.971	0.952	0.936	0.944
		grain	0.985	0.916	0.852	0.883
		interest	0.97	0.69	0.698	0.694
		money-fx	0.971	0.822	0.783	0.802
		ship	0.99	0.905	0.77	0.832
		trade	0.969	0.752	0.68	0.714
		wheat	0.991	0.849	0.86	0.855
	$score_B$	acq	0.967	0.933	0.931	0.932
		corn	0.992	0.854	0.825	0.839
		crude	0.978	0.847	0.814	0.83
		earn	0.972	0.952	0.939	0.945
		grain	0.985	0.913	0.85	0.88
		interest	0.969	0.685	0.693	0.689
		money-fx	0.97	0.821	0.783	0.801
		ship	0.99	0.91	0.77	0.834
		trade	0.969	0.752	0.68	0.714
		wheat	0.991	0.853	0.86	0.857
	$score_C$	acq	0.967	0.934	0.931	0.933
		corn	0.992	0.85	0.831	0.84
		crude	0.978	0.844	0.816	0.83
		earn	0.971	0.952	0.936	0.944
		grain	0.985	0.911	0.847	0.878
		interest	0.97	0.693	0.693	0.693
		money-fx	0.971	0.824	0.784	0.804
		ship	0.99	0.914	0.77	0.836
		trade	0.969	0.748	0.68	0.712
		wheat	0.991	0.863	0.856	0.86

表 A.8: 高頻度かつ NN で素性選択されたユニグラム＋共起単語の実験結果 (素性数が 20000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
20000	$score_A$	acq	0.967	0.933	0.933	0.933
		corn	0.992	0.855	0.831	0.842
		crude	0.978	0.843	0.814	0.829
		earn	0.971	0.95	0.936	0.943
		grain	0.985	0.913	0.845	0.878
		interest	0.97	0.69	0.698	0.694
		money-fx	0.971	0.823	0.786	0.804
		ship	0.99	0.905	0.77	0.832
		trade	0.969	0.752	0.68	0.714
		wheat	0.991	0.849	0.86	0.855
	$score_B$	acq	0.967	0.934	0.931	0.933
		corn	0.992	0.853	0.819	0.836
		crude	0.979	0.851	0.816	0.834
		earn	0.972	0.953	0.937	0.945
		grain	0.986	0.916	0.856	0.885
		interest	0.969	0.683	0.693	0.688
		money-fx	0.97	0.819	0.784	0.802
		ship	0.99	0.91	0.77	0.834
		trade	0.97	0.758	0.678	0.715
		wheat	0.991	0.853	0.86	0.857
	$score_C$	acq	0.967	0.934	0.933	0.934
		corn	0.992	0.855	0.831	0.842
		crude	0.979	0.85	0.823	0.836
		earn	0.971	0.952	0.935	0.944
		grain	0.985	0.913	0.843	0.877
		interest	0.969	0.685	0.693	0.689
		money-fx	0.971	0.823	0.786	0.804
		ship	0.99	0.915	0.779	0.841
		trade	0.969	0.748	0.688	0.717
		wheat	0.991	0.853	0.86	0.857

表 A.9: 高頻度かつ NN で素性選択されたユニグラム＋共起単語の実験結果 (素性数が 25000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
25000	$score_A$	acq	0.967	0.934	0.932	0.933
		corn	0.992	0.854	0.825	0.839
		crude	0.978	0.845	0.814	0.829
		earn	0.971	0.95	0.937	0.944
		grain	0.985	0.911	0.847	0.878
		interest	0.969	0.687	0.693	0.69
		money-fx	0.971	0.82	0.79	0.805
		ship	0.99	0.905	0.766	0.829
		trade	0.97	0.754	0.68	0.715
		wheat	0.991	0.849	0.86	0.855
	$score_B$	acq	0.967	0.933	0.933	0.933
		corn	0.992	0.853	0.819	0.836
		crude	0.979	0.852	0.823	0.837
		earn	0.972	0.953	0.937	0.945
		grain	0.985	0.909	0.852	0.88
		interest	0.969	0.686	0.698	0.692
		money-fx	0.97	0.821	0.783	0.801
		ship	0.99	0.91	0.77	0.834
		trade	0.97	0.762	0.678	0.717
		wheat	0.991	0.858	0.869	0.863
	$score_C$	acq	0.967	0.933	0.932	0.932
		corn	0.992	0.854	0.825	0.839
		crude	0.979	0.849	0.823	0.836
		earn	0.971	0.951	0.935	0.943
		grain	0.985	0.913	0.847	0.879
		interest	0.969	0.688	0.69	0.689
		money-fx	0.971	0.824	0.79	0.806
		ship	0.99	0.909	0.766	0.831
		trade	0.97	0.752	0.688	0.718
		wheat	0.991	0.853	0.86	0.857

表 A.10: 高頻度のユニグラム＋素性選択された共起単語の実験結果 (N_t=25 のとき)

N_t	スコア付け	カテゴリ	正答率	精度	再現率	F 値
25	$score_A$	acq	0.962	0.915	0.932	0.923
		corn	0.983	0.854	0.413	0.557
		crude	0.978	0.887	0.757	0.817
		earn	0.967	0.942	0.925	0.934
		grain	0.979	0.911	0.746	0.82
		interest	0.969	0.69	0.61	0.648
		money-fx	0.965	0.798	0.726	0.76
		ship	0.978	0.951	0.39	0.554
		trade	0.967	0.808	0.57	0.669
		wheat	0.989	0.869	0.751	0.806
	$score_B$	acq	0.962	0.927	0.922	0.924
		corn	0.99	0.824	0.733	0.775
		crude	0.979	0.874	0.775	0.821
		earn	0.971	0.944	0.94	0.942
		grain	0.98	0.921	0.752	0.828
		interest	0.967	0.679	0.603	0.639
		money-fx	0.966	0.825	0.731	0.775
		ship	0.985	0.838	0.655	0.735
		trade	0.965	0.741	0.615	0.672
		wheat	0.988	0.843	0.764	0.802
	$score_C$	acq	0.963	0.935	0.917	0.926
		corn	0.986	0.836	0.567	0.675
		crude	0.975	0.875	0.717	0.789
		earn	0.967	0.943	0.926	0.935
		grain	0.982	0.885	0.815	0.849
		interest	0.97	0.775	0.56	0.65
		money-fx	0.963	0.764	0.756	0.76
		ship	0.981	0.915	0.52	0.663
		trade	0.965	0.736	0.588	0.654
		wheat	0.989	0.86	0.795	0.826

表 A.11: 高頻度のユニグラム＋素性選択された共起単語の実験結果 (N_t=50 のとき)

N_t	スコア付け	カテゴリ	正答率	精度	再現率	F 値
50	$score_A$	acq	0.957	0.904	0.926	0.915
		corn	0.985	0.86	0.478	0.614
		crude	0.977	0.882	0.741	0.805
		earn	0.969	0.942	0.936	0.939
		grain	0.98	0.921	0.76	0.833
		interest	0.968	0.678	0.593	0.633
		money-fx	0.963	0.792	0.709	0.748
		ship	0.985	0.852	0.624	0.721
		trade	0.962	0.714	0.558	0.626
		wheat	0.991	0.848	0.859	0.853
	$score_B$	acq	0.962	0.931	0.918	0.924
		corn	0.982	0.82	0.392	0.531
		crude	0.978	0.876	0.77	0.82
		earn	0.971	0.95	0.938	0.944
		grain	0.979	0.869	0.801	0.834
		interest	0.971	0.763	0.59	0.666
		money-fx	0.966	0.807	0.759	0.783
		ship	0.983	0.815	0.595	0.688
		trade	0.965	0.76	0.589	0.664
		wheat	0.989	0.822	0.822	0.822
	$score_C$	acq	0.956	0.908	0.91	0.909
		corn	0.984	0.848	0.467	0.602
		crude	0.975	0.861	0.721	0.785
		earn	0.969	0.965	0.914	0.939
		grain	0.982	0.909	0.806	0.855
		interest	0.967	0.686	0.572	0.624
		money-fx	0.968	0.832	0.735	0.781
		ship	0.982	0.872	0.551	0.675
		trade	0.96	0.694	0.554	0.616
		wheat	0.992	0.871	0.855	0.863

表 A.12: 高頻度のユニグラム＋素性選択された共起単語の実験結果 (N_t=100 のとき)

N_t	スコア付け	カテゴリ	正答率	精度	再現率	F 値
100	$score_A$	acq	0.959	0.916	0.921	0.919
		corn	0.985	0.76	0.548	0.637
		crude	0.977	0.875	0.742	0.803
		earn	0.97	0.951	0.931	0.941
		grain	0.976	0.877	0.737	0.801
		interest	0.97	0.693	0.615	0.652
		money-fx	0.971	0.839	0.774	0.805
		ship	0.983	0.885	0.544	0.674
		trade	0.965	0.775	0.577	0.661
		wheat	0.987	0.825	0.768	0.796
	$score_B$	acq	0.96	0.904	0.936	0.919
		corn	0.987	0.815	0.601	0.692
		crude	0.976	0.866	0.741	0.799
		earn	0.966	0.965	0.901	0.932
		grain	0.981	0.917	0.774	0.839
		interest	0.969	0.68	0.65	0.665
		money-fx	0.966	0.792	0.745	0.768
		ship	0.983	0.872	0.539	0.667
		trade	0.965	0.781	0.563	0.654
		wheat	0.987	0.831	0.771	0.8
	$score_C$	acq	0.955	0.923	0.896	0.909
		corn	0.985	0.853	0.486	0.619
		crude	0.978	0.875	0.753	0.81
		earn	0.966	0.965	0.901	0.932
		grain	0.98	0.907	0.771	0.834
		interest	0.97	0.7	0.631	0.664
		money-fx	0.969	0.806	0.787	0.797
		ship	0.984	0.911	0.568	0.7
		trade	0.964	0.736	0.593	0.657
		wheat	0.989	0.871	0.772	0.819

表 A.13: 高頻度のユニグラム＋素性選択された共起単語の実験結果 (N_t=125 のとき)

N_t	スコア付け	カテゴリ	正答率	精度	再現率	F 値
125	$score_A$	acq	0.962	0.924	0.924	0.924
		corn	0.989	0.816	0.697	0.752
		crude	0.976	0.839	0.749	0.792
		earn	0.964	0.946	0.912	0.929
		grain	0.979	0.935	0.733	0.822
		interest	0.964	0.696	0.496	0.579
		money-fx	0.966	0.791	0.744	0.767
		ship	0.982	0.895	0.54	0.674
		trade	0.964	0.721	0.599	0.654
		wheat	0.989	0.849	0.793	0.82
	$score_B$	acq	0.963	0.927	0.924	0.926
		corn	0.989	0.897	0.631	0.741
		crude	0.978	0.856	0.77	0.811
		earn	0.969	0.956	0.922	0.939
		grain	0.979	0.91	0.753	0.824
		interest	0.966	0.71	0.548	0.619
		money-fx	0.964	0.834	0.667	0.741
		ship	0.982	0.88	0.53	0.661
		trade	0.967	0.755	0.639	0.692
		wheat	0.99	0.836	0.829	0.833
	$score_C$	acq	0.962	0.916	0.932	0.924
		corn	0.985	0.847	0.477	0.61
		crude	0.978	0.886	0.737	0.805
		earn	0.971	0.963	0.922	0.942
		grain	0.976	0.866	0.743	0.8
		interest	0.969	0.715	0.637	0.674
		money-fx	0.966	0.826	0.698	0.757
		ship	0.984	0.913	0.583	0.712
		trade	0.966	0.788	0.562	0.656
		wheat	0.989	0.828	0.843	0.835

表 A.14: 高頻度のユニグラム＋素性選択された共起単語の実験結果 (N_t=150 のとき)

N_t	スコア付け	カテゴリ	正答率	精度	再現率	F 値
150	$score_A$	acq	0.963	0.925	0.926	0.925
		corn	0.988	0.827	0.621	0.709
		crude	0.974	0.875	0.698	0.777
		earn	0.962	0.942	0.908	0.924
		grain	0.979	0.891	0.768	0.825
		interest	0.969	0.713	0.617	0.662
		money-fx	0.97	0.853	0.753	0.8
		ship	0.983	0.872	0.542	0.668
		trade	0.966	0.745	0.618	0.676
		wheat	0.988	0.829	0.774	0.801
	$score_B$	acq	0.957	0.907	0.919	0.913
		corn	0.99	0.884	0.678	0.767
		crude	0.979	0.865	0.796	0.829
		earn	0.967	0.94	0.929	0.934
		grain	0.98	0.897	0.774	0.831
		interest	0.971	0.718	0.664	0.69
		money-fx	0.969	0.823	0.763	0.792
		ship	0.98	0.879	0.498	0.636
		trade	0.966	0.787	0.571	0.662
		wheat	0.987	0.791	0.783	0.787
	$score_C$	acq	0.963	0.937	0.914	0.926
		corn	0.985	0.833	0.508	0.632
		crude	0.974	0.886	0.696	0.779
		earn	0.965	0.954	0.904	0.928
		grain	0.979	0.889	0.766	0.823
		interest	0.97	0.704	0.644	0.673
		money-fx	0.969	0.826	0.761	0.792
		ship	0.984	0.868	0.595	0.706
		trade	0.962	0.713	0.563	0.629
		wheat	0.99	0.83	0.86	0.844

表 A.15: 出現頻度による分割学習の結果 (素性集合が高頻度かつ NN で素性選択されたユニグラムのとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
2500	$score_A$	acq	0.926	0.825	0.875	0.849
		earn	0.919	0.821	0.851	0.836
		trade	0.95	0.576	0.565	0.57
	$score_B$	acq	0.944	0.887	0.887	0.887
		earn	0.958	0.917	0.914	0.915
		trade	0.962	0.676	0.694	0.685
	$score_C$	acq	0.928	0.841	0.875	0.858
		earn	0.918	0.792	0.816	0.804
		trade	0.952	0.6	0.574	0.587
5000	$score_A$	acq	0.931	0.841	0.884	0.862
		earn	0.933	0.853	0.884	0.868
		trade	0.951	0.571	0.588	0.579
	$score_B$	acq	0.944	0.888	0.886	0.887
		earn	0.962	0.927	0.918	0.923
		trade	0.966	0.718	0.708	0.713
	$score_C$	acq	0.938	0.86	0.9	0.88
		earn	0.925	0.813	0.857	0.834
		trade	0.953	0.583	0.61	0.596

表 A.16: NN で素性選択されたユニグラムの実験結果 (素性数が 3000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
3000	$score_A$	acq	0.859	0.744	0.65	0.694
		corn	0.981	0.68	0.4	0.504
		crude	0.951	0.711	0.448	0.55
		earn	0.917	0.883	0.776	0.826
		grain	0.939	0.571	0.21	0.307
		interest	0.959	0.614	0.37	0.462
		money-fx	0.938	0.687	0.401	0.506
		ship	0.975	0.765	0.292	0.423
		trade	0.951	0.625	0.463	0.532
		wheat	0.962	0.404	0.091	0.149
	$score_B$	acq	0.884	0.822	0.696	0.754
		corn	0.972	0.449	0.124	0.194
		crude	0.953	0.745	0.474	0.579
		earn	0.92	0.901	0.778	0.835
		grain	0.953	0.683	0.508	0.582
		interest	0.953	0.532	0.261	0.35
		money-fx	0.947	0.705	0.541	0.612
		ship	0.972	0.739	0.305	0.432
		trade	0.955	0.667	0.444	0.533
		wheat	0.971	0.583	0.275	0.374
	$score_C$	acq	0.858	0.747	0.652	0.697
		corn	0.978	0.67	0.379	0.484
		crude	0.95	0.722	0.413	0.525
		earn	0.913	0.883	0.769	0.822
		grain	0.94	0.6	0.191	0.29
		interest	0.962	0.615	0.347	0.443
		money-fx	0.936	0.639	0.395	0.488
		ship	0.971	0.667	0.219	0.33
		trade	0.947	0.64	0.42	0.507
		wheat	0.969	0.629	0.291	0.398

表 A.17: NN で素性選択されたユニグラムの実験結果 (素性数が 6000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
6000	$score_A$	acq	0.888	0.79	0.758	0.774
		corn	0.976	0.545	0.436	0.484
		crude	0.961	0.732	0.608	0.664
		earn	0.935	0.895	0.845	0.87
		grain	0.956	0.69	0.56	0.618
		interest	0.963	0.618	0.529	0.57
		money-fx	0.949	0.688	0.607	0.645
		ship	0.976	0.675	0.48	0.561
		trade	0.956	0.642	0.564	0.601
		wheat	0.967	0.466	0.376	0.416
	$score_B$	acq	0.92	0.858	0.817	0.837
		corn	0.968	0.355	0.303	0.327
		crude	0.957	0.701	0.562	0.624
		earn	0.943	0.9	0.876	0.888
		grain	0.959	0.685	0.659	0.672
		interest	0.962	0.615	0.546	0.578
		money-fx	0.956	0.733	0.696	0.714
		ship	0.974	0.649	0.489	0.558
		trade	0.963	0.689	0.629	0.658
		wheat	0.968	0.519	0.41	0.459
	$score_C$	acq	0.895	0.796	0.772	0.783
		corn	0.976	0.542	0.433	0.481
		crude	0.962	0.734	0.613	0.668
		earn	0.922	0.875	0.808	0.84
		grain	0.95	0.629	0.545	0.584
		interest	0.961	0.611	0.525	0.565
		money-fx	0.955	0.721	0.661	0.69
		ship	0.975	0.673	0.437	0.53
		trade	0.958	0.657	0.576	0.614
		wheat	0.972	0.551	0.452	0.496

表 A.18: NN で素性選択されたユニグラムの実験結果 (素性数が 9000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
9000	$score_A$	acq	0.917	0.837	0.833	0.835
		corn	0.99	0.797	0.828	0.812
		crude	0.958	0.695	0.631	0.661
		earn	0.938	0.888	0.859	0.874
		grain	0.966	0.751	0.734	0.742
		interest	0.962	0.597	0.583	0.59
		money-fx	0.964	0.759	0.777	0.768
		ship	0.978	0.715	0.522	0.604
		trade	0.957	0.645	0.617	0.631
		wheat	0.973	0.559	0.528	0.543
	$score_B$	acq	0.928	0.865	0.844	0.854
		corn	0.968	0.392	0.346	0.367
		crude	0.958	0.678	0.629	0.653
		earn	0.956	0.915	0.908	0.911
		grain	0.97	0.775	0.729	0.751
		interest	0.965	0.649	0.632	0.64
		money-fx	0.958	0.726	0.731	0.729
		ship	0.98	0.737	0.606	0.665
		trade	0.96	0.662	0.63	0.646
		wheat	0.967	0.506	0.494	0.5
	$score_C$	acq	0.925	0.849	0.843	0.846
		corn	0.979	0.567	0.515	0.54
		crude	0.962	0.713	0.624	0.666
		earn	0.939	0.891	0.864	0.877
		grain	0.955	0.664	0.646	0.655
		interest	0.961	0.601	0.572	0.586
		money-fx	0.952	0.702	0.691	0.696
		ship	0.978	0.707	0.549	0.618
		trade	0.966	0.711	0.684	0.697
		wheat	0.967	0.496	0.5	0.498

表 A.19: NN で素性選択されたユニグラムの実験結果 (素性数が 12000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
12000	$score_A$	acq	0.929	0.862	0.855	0.858
		corn	0.99	0.8	0.8	0.8
		crude	0.963	0.732	0.681	0.706
		earn	0.945	0.898	0.883	0.89
		grain	0.976	0.809	0.812	0.811
		interest	0.968	0.637	0.651	0.644
		money-fx	0.962	0.753	0.753	0.753
		ship	0.98	0.746	0.603	0.667
		trade	0.96	0.665	0.621	0.642
		wheat	0.968	0.489	0.493	0.491
	$score_B$	acq	0.937	0.878	0.869	0.873
		corn	0.971	0.431	0.431	0.431
		crude	0.975	0.803	0.803	0.803
		earn	0.953	0.906	0.904	0.905
		grain	0.968	0.762	0.747	0.755
		interest	0.966	0.645	0.659	0.652
		money-fx	0.959	0.722	0.74	0.731
		ship	0.982	0.765	0.695	0.728
		trade	0.965	0.717	0.668	0.692
		wheat	0.991	0.847	0.866	0.857
	$score_C$	acq	0.931	0.864	0.852	0.858
		corn	0.991	0.821	0.825	0.823
		crude	0.972	0.794	0.726	0.758
		earn	0.947	0.904	0.887	0.895
		grain	0.961	0.713	0.701	0.707
		interest	0.968	0.649	0.677	0.663
		money-fx	0.954	0.7	0.712	0.706
		ship	0.979	0.705	0.594	0.645
		trade	0.967	0.724	0.719	0.722
		wheat	0.971	0.544	0.57	0.557

表 A.20: NN で素性選択されたユニグラムの実験結果 (素性数が 15000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
15000	$score_A$	acq	0.945	0.893	0.887	0.89
		corn	0.991	0.806	0.829	0.817
		crude	0.969	0.779	0.722	0.749
		earn	0.945	0.894	0.887	0.891
		grain	0.976	0.83	0.81	0.82
		interest	0.965	0.632	0.626	0.629
		money-fx	0.963	0.758	0.761	0.759
		ship	0.984	0.824	0.675	0.742
		trade	0.965	0.707	0.695	0.701
		wheat	0.967	0.523	0.556	0.539
	$score_B$	acq	0.948	0.902	0.892	0.897
		corn	0.98	0.605	0.514	0.556
		crude	0.978	0.837	0.813	0.824
		earn	0.963	0.932	0.922	0.927
		grain	0.98	0.851	0.819	0.835
		interest	0.97	0.693	0.674	0.683
		money-fx	0.962	0.746	0.758	0.752
		ship	0.986	0.83	0.706	0.763
		trade	0.97	0.744	0.731	0.737
		wheat	0.989	0.833	0.833	0.833
	$score_C$	acq	0.94	0.88	0.879	0.88
		corn	0.991	0.801	0.806	0.804
		crude	0.975	0.812	0.763	0.787
		earn	0.953	0.908	0.909	0.908
		grain	0.973	0.784	0.784	0.784
		interest	0.968	0.653	0.671	0.662
		money-fx	0.963	0.766	0.766	0.766
		ship	0.983	0.8	0.658	0.722
		trade	0.969	0.733	0.724	0.728
		wheat	0.971	0.561	0.546	0.553

表 A.21: NN で素性選択されたユニグラムの実験結果 (素性数が 18000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
18000	$score_A$	acq	0.95	0.902	0.897	0.9
		corn	0.992	0.813	0.841	0.827
		crude	0.972	0.819	0.715	0.763
		earn	0.956	0.919	0.907	0.913
		grain	0.978	0.858	0.814	0.835
		interest	0.967	0.64	0.691	0.665
		money-fx	0.968	0.796	0.765	0.78
		ship	0.984	0.825	0.679	0.745
		trade	0.97	0.739	0.73	0.734
		wheat	0.989	0.827	0.844	0.835
	$score_B$	acq	0.955	0.912	0.907	0.91
		corn	0.993	0.828	0.89	0.858
		crude	0.981	0.876	0.821	0.848
		earn	0.965	0.933	0.928	0.931
		grain	0.986	0.898	0.88	0.889
		interest	0.969	0.688	0.672	0.679
		money-fx	0.963	0.759	0.757	0.758
		ship	0.986	0.841	0.697	0.763
		trade	0.971	0.756	0.741	0.748
		wheat	0.99	0.847	0.861	0.854
	$score_C$	acq	0.951	0.901	0.903	0.902
		corn	0.992	0.828	0.846	0.837
		crude	0.974	0.807	0.795	0.801
		earn	0.952	0.904	0.905	0.904
		grain	0.974	0.8	0.795	0.798
		interest	0.97	0.686	0.688	0.687
		money-fx	0.966	0.786	0.779	0.782
		ship	0.983	0.802	0.683	0.738
		trade	0.966	0.713	0.701	0.707
		wheat	0.973	0.584	0.559	0.571

表 A.22: NN で素性選択されたユニグラムの実験結果 (素性数が 21000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
21000	$score_A$	acq	0.96	0.924	0.918	0.921
		corn	0.993	0.84	0.855	0.847
		crude	0.972	0.804	0.716	0.758
		earn	0.968	0.942	0.93	0.936
		grain	0.977	0.839	0.787	0.812
		interest	0.969	0.676	0.674	0.675
		money-fx	0.968	0.793	0.772	0.782
		ship	0.988	0.868	0.734	0.795
		trade	0.966	0.699	0.707	0.703
		wheat	0.99	0.85	0.828	0.838
	$score_B$	acq	0.959	0.913	0.922	0.918
		corn	0.99	0.798	0.825	0.811
		crude	0.98	0.863	0.816	0.839
		earn	0.968	0.942	0.929	0.936
		grain	0.987	0.901	0.884	0.892
		interest	0.967	0.658	0.687	0.672
		money-fx	0.968	0.79	0.797	0.794
		ship	0.989	0.891	0.762	0.821
		trade	0.972	0.758	0.766	0.762
		wheat	0.991	0.865	0.869	0.867
	$score_C$	acq	0.958	0.918	0.915	0.917
		corn	0.992	0.829	0.852	0.84
		crude	0.977	0.833	0.81	0.821
		earn	0.962	0.93	0.922	0.926
		grain	0.97	0.77	0.768	0.769
		interest	0.972	0.71	0.722	0.716
		money-fx	0.967	0.789	0.761	0.775
		ship	0.986	0.856	0.687	0.762
		trade	0.967	0.734	0.696	0.714
		wheat	0.99	0.855	0.847	0.851

表 A.23: NN で素性選択されたユニグラムの実験結果 (素性数が 24000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
24000	$score_A$	acq	0.96	0.915	0.922	0.918
		corn	0.992	0.82	0.849	0.834
		crude	0.974	0.853	0.737	0.79
		earn	0.969	0.944	0.937	0.94
		grain	0.979	0.866	0.8	0.831
		interest	0.967	0.666	0.638	0.652
		money-fx	0.968	0.8	0.761	0.78
		ship	0.989	0.898	0.76	0.823
		trade	0.97	0.739	0.708	0.723
		wheat	0.989	0.848	0.819	0.833
	$score_B$	acq	0.965	0.929	0.933	0.931
		corn	0.991	0.83	0.825	0.827
		crude	0.98	0.864	0.821	0.842
		earn	0.969	0.947	0.932	0.939
		grain	0.986	0.908	0.865	0.886
		interest	0.967	0.662	0.675	0.669
		money-fx	0.967	0.793	0.77	0.781
		ship	0.989	0.892	0.77	0.826
		trade	0.97	0.749	0.712	0.73
		wheat	0.99	0.858	0.843	0.85
	$score_C$	acq	0.963	0.924	0.924	0.924
		corn	0.994	0.862	0.881	0.872
		crude	0.979	0.858	0.821	0.839
		earn	0.969	0.942	0.936	0.939
		grain	0.982	0.879	0.848	0.863
		interest	0.97	0.689	0.705	0.697
		money-fx	0.968	0.794	0.769	0.781
		ship	0.989	0.888	0.751	0.814
		trade	0.969	0.751	0.711	0.73
		wheat	0.99	0.843	0.847	0.845

表 A.24: NN で素性選択されたユニグラムの実験結果 (素性数が 27000 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
27000	$score_A$	acq	0.966	0.934	0.929	0.931
		corn	0.992	0.834	0.853	0.844
		crude	0.979	0.861	0.816	0.838
		earn	0.969	0.946	0.935	0.94
		grain	0.983	0.891	0.841	0.865
		interest	0.969	0.678	0.684	0.681
		money-fx	0.969	0.808	0.784	0.796
		ship	0.989	0.875	0.774	0.822
		trade	0.971	0.759	0.72	0.739
		wheat	0.991	0.86	0.86	0.86
	$score_B$	acq	0.97	0.94	0.938	0.939
		corn	0.992	0.839	0.825	0.832
		crude	0.978	0.846	0.808	0.827
		earn	0.974	0.957	0.939	0.948
		grain	0.984	0.898	0.843	0.87
		interest	0.968	0.669	0.687	0.678
		money-fx	0.97	0.816	0.79	0.803
		ship	0.989	0.871	0.774	0.82
		trade	0.971	0.749	0.72	0.734
		wheat	0.99	0.852	0.852	0.852
	$score_C$	acq	0.966	0.93	0.931	0.931
		corn	0.993	0.853	0.853	0.853
		crude	0.98	0.863	0.827	0.845
		earn	0.969	0.946	0.933	0.94
		grain	0.985	0.91	0.858	0.883
		interest	0.97	0.693	0.705	0.699
		money-fx	0.971	0.813	0.807	0.81
		ship	0.989	0.884	0.755	0.815
		trade	0.972	0.763	0.715	0.739
		wheat	0.99	0.847	0.847	0.847

表 A.25: NN で素性選択されたユニグラムの実験結果 (素性数が 28795 個のとき)

素性数	スコア付け	カテゴリ	正答率	精度	再現率	F 値
28795	$score_A$	acq	0.971	0.942	0.941	0.942
		corn	0.993	0.852	0.847	0.85
		crude	0.978	0.852	0.808	0.829
		earn	0.972	0.953	0.936	0.945
		grain	0.985	0.91	0.861	0.885
		interest	0.968	0.673	0.67	0.671
		money-fx	0.972	0.818	0.81	0.814
		ship	0.989	0.864	0.783	0.821
		trade	0.973	0.784	0.705	0.743
		wheat	0.991	0.853	0.86	0.857
	$score_B$	acq	0.969	0.936	0.938	0.937
		corn	0.992	0.834	0.825	0.83
		crude	0.978	0.856	0.806	0.83
		earn	0.972	0.956	0.936	0.946
		grain	0.986	0.91	0.863	0.886
		interest	0.968	0.665	0.695	0.68
		money-fx	0.975	0.842	0.823	0.833
		ship	0.988	0.888	0.74	0.807
		trade	0.973	0.773	0.73	0.751
		wheat	0.991	0.854	0.869	0.861
	$score_C$	acq	0.969	0.936	0.937	0.937
		corn	0.992	0.84	0.859	0.849
		crude	0.98	0.863	0.816	0.839
		earn	0.972	0.954	0.937	0.945
		grain	0.985	0.908	0.863	0.885
		interest	0.969	0.68	0.684	0.682
		money-fx	0.972	0.825	0.805	0.815
		ship	0.991	0.92	0.783	0.846
		trade	0.972	0.774	0.718	0.745
		wheat	0.991	0.84	0.873	0.857