第３章　景気ウォッチャー調査の個票と各種経済指標の関係に関する検証の結果

３－１　分析作業の内容と検証手順の概観

内閣府が作成･公表した 2010 年 1 月から 2013 年 12 月の景気ウォッチャー調査の個票と各種経済指標の関係変化などを、ビッグデータ業務で用いるテキスト分析や回帰分析などの手法によって調査・分析する。これにより、東日本大震災経験後の我が国において経済構造変化や景気循環変化が起こったかどうか、並びに景況感の変化を示す統計数値（DI 値等）が景気ウォッチャー調査のコメントによって推計が可能かどうか等についての検証を試みる。

（東日本大震災に関わる分析については３－４で記述する）

景気ウォッチャー調査のコメントによる景況感の変化を示す統計数値の推計については、具体的には、景気ウォッチャー調査のコメントについては類似の意味をもつことばを統一化した上で数値化し、景気ウォッチャー調査の現状判断 DI、先行き判断 DI、地域別支出総合指数とその構成要素（地域別民間住宅総合指数、地域別公共投資総合指数等）等との関係性を検証する。検証方法は単純な回帰式など簡便な方法のみならず、日本電気株式会社（以下、ＮＥＣ）の中央研究所において開発した、特徴表現抽出などの自然言語処理技術を用いたテキスト分析、多種多様なデータから複数の規則性を容易に導出できる異種混合学習技術を活用したアプローチを試みる。

３－１－１　全体フロー

分析の全体フローの概観を以下に示す。

分析作業を２つに大別すると、テキストマイニング（下記①～⑦）とデータマイニング（下記⑧～⑩）に分けられる。

＜①～⑦：テキストマイニング作業　３－２参照＞

① **コメントからの単語の抽出：**景気ウォッチャー調査のコメントから、単語を抽出する。単語は、名詞や動詞等の自立語のみを選択し、助詞等の付属語は削除する。また、動詞などの活用によって表現が異なる単語を同一視するために、抽出した単語は原形に変換する。

② **景気に関連する特徴表現の抽出：**景気ウォッチャー調査のコメントから、景気動向に関連する特徴的な表現（単語、フレーズ）を抽出する。

③ **同義語の抽出：**複数の同義判定方式を用いて、景気ウォッチャー調査のコメント中から同義語を抽出する。抽出した同義語を元に、同じ意味の単語を同義語中の

代表語で纏め上げる。

④ **有意表現の抽出:**纏め上げられた単語と特徴表現の中から景気判断(「よくなった」、「悪くなった」)の判別に影響を与える上位 400 語を有意表現として抽出する。

⑤ **コメントからの景気現状判断分類（全国、地域別）:** 景気ウォッチャー調査のコメント（テキストデータ）を、抽出した単語・特徴表現にもとづいて機械学習の判別エンジンにかけて、景気の現状判断を「良くなっている、やや良くなっている」か「悪くなっている、やや悪くなっている」かの 2 値に分類する。

⑥ **コメントからの DI 値の予測（現状判断 DI、先行き判断 DI）:** ⑤の分類結果を元に、今度はそれぞれのコメントを a)「良くなった」か「悪くなった」のいずれか（2 値)、b)「 良くなった」か「変わらない」か「悪くなった」の 3 値、c) 「良くなった」か「やや良くなったか」か「変わらない」か「やや悪くなった」か「悪くなった」の 5 値に分類し、その結果をもとに DI 値を予測する。また同様の予測作業を、先行き判断 DI 値についても行う。

⑦ **コメントからの DI 以外の経済指標の上昇／下降の予測:** テキスト情報(景気ウォッチャー調査のコメント)から、景気ウォッチャー調査 DI 以外の経済指標の上昇／下降が予測可能かを検証する。

＜⑧～⑩：データマイニング作業　３－３、３－４参照＞

⑧ **コメント（テキストデータ）の数値化**：景気指標の値そのものをテキストデータから予測、計算できないかを分析するため、有意単語 400 語の出現回数を用いてテキストデータを数値化する。

⑨ **予測式の抽出:** 数値化したデータと実際の現状判断 DI 値との関連を求める複数の予測式（回帰式）を、回帰分析ツールを使って学習区間において抽出する。

⑩ **予測式の評価:** 学習区間で自動抽出した複数の予測式を、予測区間にあてはめて、同じ関係性が成立しているかをチェックする。

⑤、⑦は有意単語ではなく全単語を用い、⑤、⑦、⑧については並行して実施できるので、フローとしては以下のようになる。

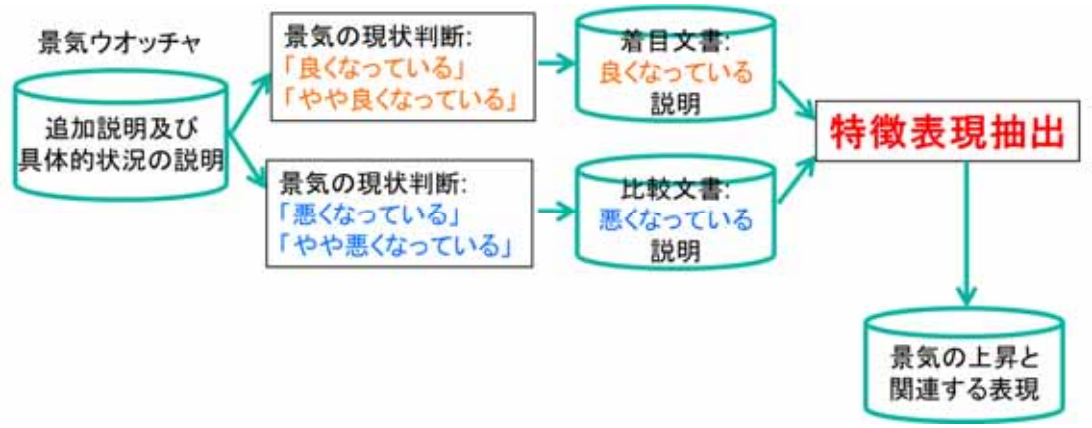
①　⇒②　⇒③　⇒⑤　⇒⑥
　　⇓　　⑦
　　④　⇒⑧　⇒⑨　⇒⑩

なお、３－３においては分析の対象／目的別に分析Ａ、Ｂ、Ｃ、Ｄの４つ、３－４においては分析Ｅ、Ｆの２つを実施している。(それぞれの分析概要を記した比較表は３－３、３－４を参照)

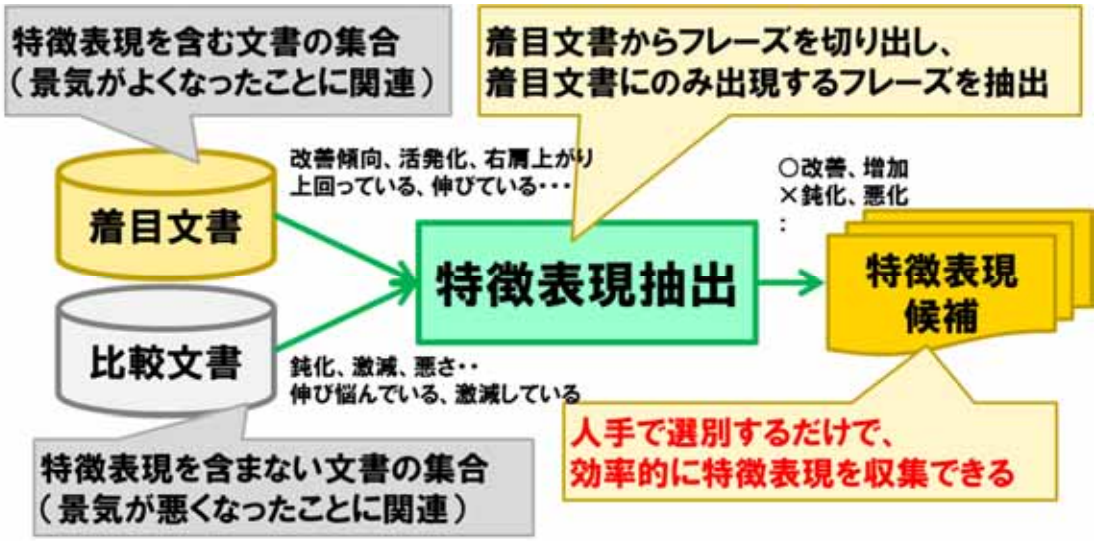
以下、上記①～⑩のうち、技術面の詳細説明が必要な作業項目につき解説する。

３－１－２　景気に関連する特徴表現の抽出（上記②）

景気ウォッチャー調査の「追加説明及び具体的状況の説明」のコメント（以下、コメント）から景気の動向に関連する特徴表現を抽出する。抽出方法は、まずコメントのテキストデータから、景気動向が「良くなった」、「やや良くなった」に対応する部分のみを選択し、着目文書という集合を作成する。次に景気動向が「悪くなった」、「やや悪くなった」に対応する部分のみを選択し、比較文書という集合を作成する。



着目文書と比較文書を元に特徴表現の候補が自動抽出される。抽出にあたっては、ＮＥＣが開発した特徴表現抽出ツールを用いており、単語の特徴語を抽出するだけでなく、単語の並びがフレーズになり易いかを考慮することで、フレーズの特徴表現も抽出している[1]。

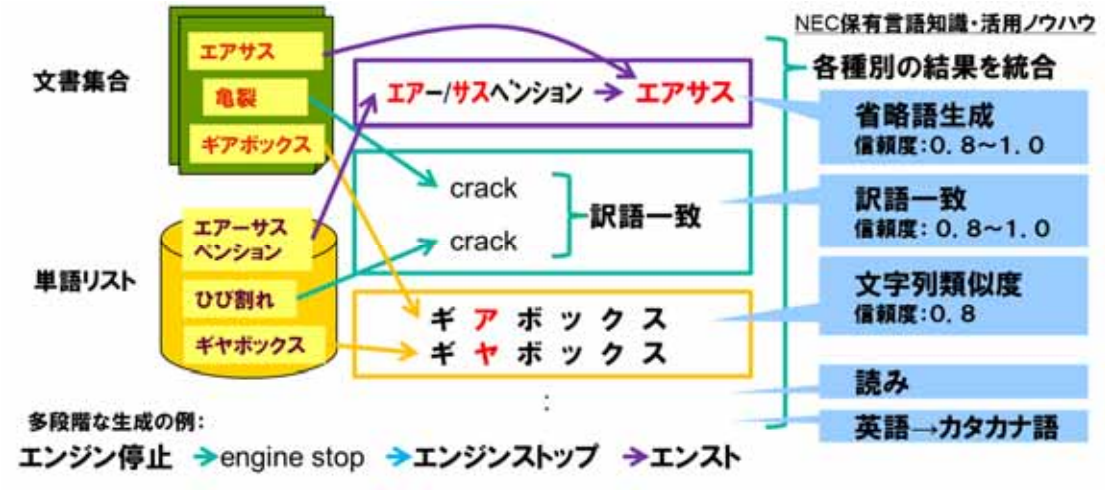


[1] また、特徴表現抽出ツールを利用することで、手作業による抽出に比べて辞書の構築工数を大幅に削減することが可能で、別タスクでの評価では約６０％削減を確認している。

同様の手法で東日本大震災および消費税に関連する特徴表現も抽出した。

３－１－３　同義語の抽出（上記③）

同義語の抽出に際し、本分析では同義語を自動抽出する技術を採用している。この技術では複数の同義語を同時に判定し、多様な同義語の候補を生成することができる。同義語には省略語（例：エアー／サスペンション→エアサス）、訳語一致（例：亀裂→クラック←ひび割れ）、文字列類似（例：ギアボックス＝ギヤボックス）、多段階生成（例：エンジン停止→エンジンストップ→エンスト）など様々なケースがあるが、これらトータルに組み合わせて同意語の候補を同時に取り出している。こうした同義語の候補には間違いもあるため、最終的には人によるチェックを行っている。



３－１－４　景気の状況判断によるコメントの分類（上記⑤）

特徴表現、同義語、有意単語の抽出をもとに、景気ウォッチャー調査のコメントを景気の状況判断にもとづいて分類（判別）する。今回分析では、コメントに現れる単語とフレーズをもとに、景気が「良い」か「悪い」のどちらか（2 値）に分類した。
分類はテキストマイニング[1]という機械処理で行う。

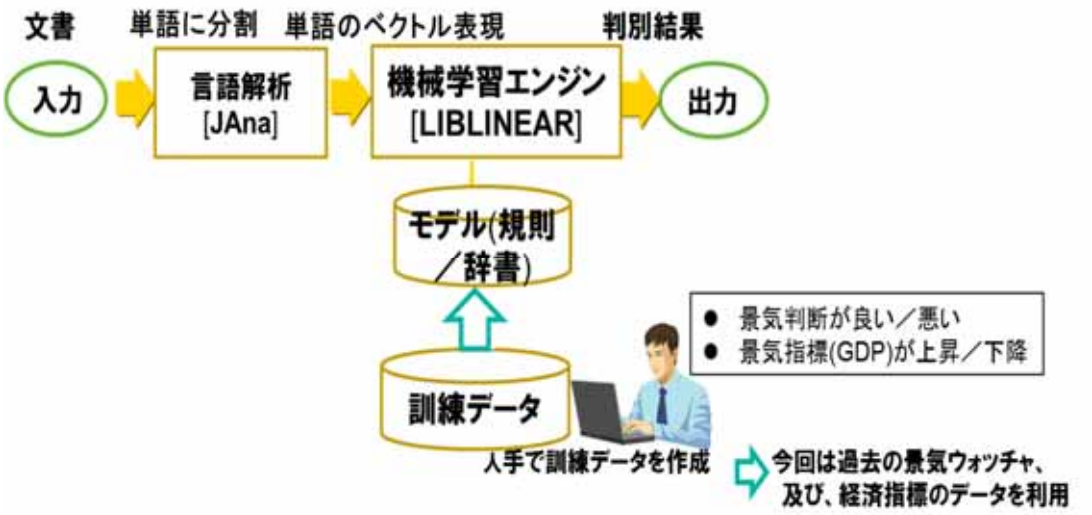
機械処理においては、機械が自動学習しながら判別精度を高める方式を用いている。はじめに正解データを用意し、正解例をもとに機械が学習してモデルを構築し、そのモデルにもとづいてコメントのテキストデータを単語分割し、単語を座標にプロットする（ベクトル表現）。コメントを 1 件ずつモデルと比較し、近いものを正解として判別していく。

[1]テキストマイニング：通常の文書からなるデータを単語や文節で区切り、それらの出現の頻度や共出現（共起ともいう）の相関、出現傾向、時系列などを解析することで有用な情報を取り出す、テキストデータの分析方法

今回、機械学習の判別ツールとしてオープンソースの LIBLINEAR[1] を利用している。

正解データは人が作成するのではなく、景気ウォッチャーの値、経済指標のデータをそのまま使用している。景気判断の良し悪しは、コメントの「景気の状況判断」の「良くなった」「やや良くなった」と「悪くなった」「やや悪くなった」を使用。また景気指標については、前月から指標が上昇したか下降したか、を見てそれを実際の数値で比べ正しかったかどうかで評価した。

機械学習ベースの技術で判別



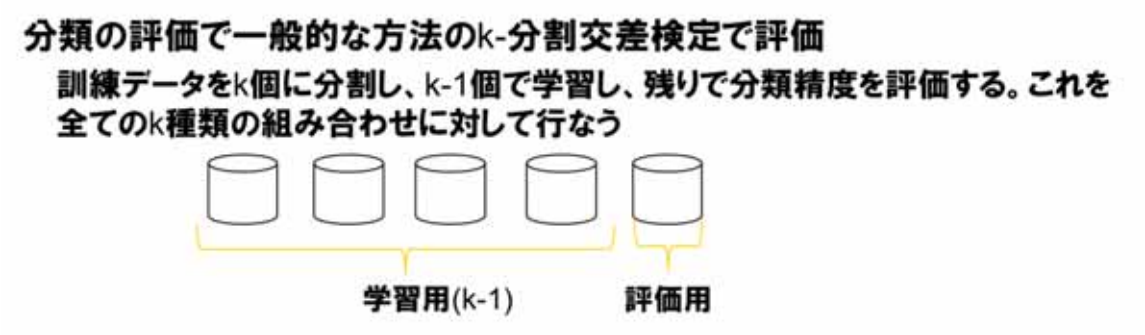
機械学習において分類（判別）の精度を高めるために今回使用した手法について説明する。

まず分類の評価手法として「k-分割交差検定 [2]」を採用している。これは、データが k 個あるとすると、ひとつの評価を残りの k－１個で学習してその分類精度を評価するもので、これを k 回繰り返すことにより、全体の分類精度を高める方式である。

---

[1]LIBLINEAR : LIBLINEAR は台湾国立大学の Chih-Jen Lin 教授のチームが公開しているオープンソースの機械学習パッケージ。C++ で書かれたライブラリと、その機能を使って機械学習と分類・関数回帰を行うコマンドラインユーティリティが含まれている

[2] 交差検定：交差検定とは、統計学においてデータを分割し、その一部をまず解析して、残る部分でその解析のテストを行い、解析自身の妥当性の検証・確認に当てる手法。



また、分類や検索でよく使用される「F 値」という手法を用いている。これは、分類で標準的に使われる尺度である適合率[1]、再現率[2]とその調和平均の F 値を用いて評価するものである。

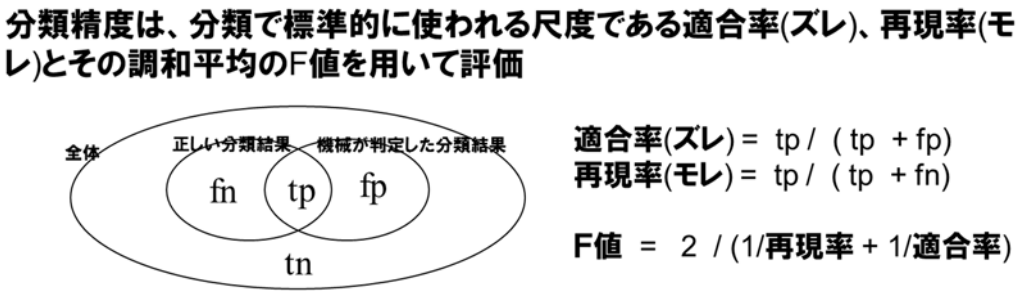
適合率と再現率はトレードオフの関係にある。一番当たりそうなものだけを選ぶと、当たる確率（適合率）は高いが、実際にはモレている可能性も高くなる。

「F 値」というのは、適合率と再現率の調和平均なので、

$$F\text{ 値} = \frac{2}{\frac{1}{\text{適合率}} + \frac{1}{\text{再現率}}}$$

となる。

F 値＝1 となるよう求めていくと分類の精度が高まる。



---

[1] 適合率:適合率とは、得られた情報の中で正解となる情報が存在する割合。正解からの「ズレ」を示す指標であり、分類されたデータのうち、どれくらい正解が含まれているか、正解でないものがどれくらい含まれているかの割合である。
100%正解なら適合率は 1 になる。50%正解なら適合率は 1／2 になる。

[2] 再現率:再現率とは、検索対象としている文書全体から正解である情報が拾い出せた割合。正解からの「モレ」を示す指標であり、本来識別されるべきものが、どのくらい識別されずにいるかを示す割合である。仮にすべて識別されたら再現率は 1 になる。半分識別できたら再現率は 1／2 になる。

３－１－５　予測式の抽出（上記⑨）

目的変数と説明変数の間の関係性モデルの抽出においては、単純な回帰分析ではなく、近年開発された異種混合学習技術にもとづく門関数付混合回帰分析を実行した。

一般的な回帰分析では、目的変数＝ f（説明変数）となる一つの回帰式を求めることしかできないが、現実のデータにおいては一つの回帰式だけでは説明ができないような場合も多く、複数の回帰式を抽出してそれを条件によってあてはめるアルゴリズムが必要になる。今回分析に使用した「異種混合学習技術」は、大量のデータから複数のパターンや法則性を導き出すことができる技術であり、同技術をアルゴリズムに用いることにより、複数の回帰式の抽出とその自動切り替えを実現している。（異種混合技術の詳細は資料９を参照）