6N-02

数値情報とテキスト情報によるバルチック海運指数の予測

滕　玲*1　　長尾 智晴*2

*1*2　横浜国立大学　大学院環境情報学府

1. 初めに

我々の身の回りには様々な時系列データが存在し，それらの予測には大きな期待が寄せられている．バルチック海運指数は時系列データで，海運市況だけでなく世界貿易の動向を占う重要な経済指標である．これまでに統計的モデル[1]やサポートベクターマシン[2]などの手法によるバルチック海運指数の予測がされており，その有効性も示されている．従来研究では，過去の指数値などの数値情報だけを用いる手法が提案されている．しかし，バルチック海運指数の変動要因は複雑であり，数値情報だけを用いた予測は困難であると考えられる．

そこで本研究では，数値情報に加えて，海運業界の関係者が参考にするニュース情報を活用することにした．数値情報とニュースのテキスト情報を用いることで，より精度の高い予測モデルを構築する．また，実験によって，提案手法の有効性を検証した．

2. 提案手法

提案手法の概要図を図 1 に示す．提案手法は（2.1）特徴量構築フェイズと（2.2）予測フェイズに分けられる．特徴量構築フェイズは数値情報の特徴量構築とテキスト情報の特徴量構築によって構成されている．予測フェイズは LSTM（Long short-term memory）ユニットを用いたニューラルネットワークを使用して数日後のバルチック海運指数の変動率の予測を行う．LSTM[3]はリカレントニューラルネットワークの一種で，時間的に長期の依存関係を学習可能なモデルである．FA.Gers ら[4]の研究によると LSTM は 1000 単位時間という長期間の性質を学習するのに成功しており，時系列データの解析における強力な学習モデルである．

2.1 特徴量構築フェイズ

数値情報は 4 種類の海運指数を使用し，過去 N 日分の指数値を数値特徴量としている．また，データの周期性を保持するため，週末を除いて週単位を 5 日間に揃える．

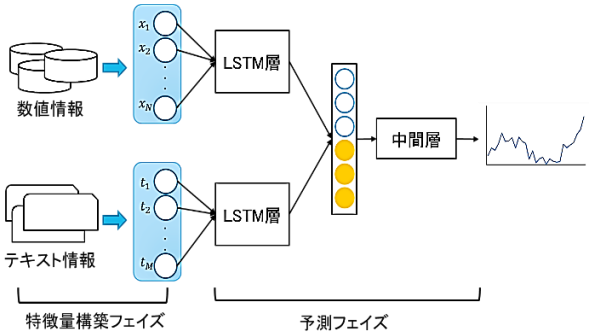

図 1:提案手法の概要

テキスト特徴量は発生した海運ニュースの影響を考慮した特徴量である．ニュースは通常，タイトルでニュースの全体像がつかめるため，本研究ではニュースのタイトルだけを使用する．その中でも特に重要なのは，これからのバルチック海運指数の変動を予測するような複合語であり，本稿ではこれを予測複合語と呼ぶ．予測複合語は 2 つ単語を仮定し，上昇（下落）の関係を持つと仮定している．GA（Genetic algorithm）を用いて，予測複合語の抽出を行った．各タイトルにおけるそれぞれの予測複合語の頻度を求めて，一定次元数のベクトル表現に変換する．日単位でテキスト情報の特徴量を構築する．予測フェイズでは，過去 M 日分のテキスト情報の特徴量を使用する．

2.2 予測フェイズ

特徴量構築フェイズから得られた特徴量を入力とし，T 日後のバルチック海運指数の変動率を出力とするニューラルネットワークの学習を行う．入力の各特徴量は[0,1]に正規化を行う．可変長の入力を考慮し，数値特徴量とテキスト特徴量をそれぞれ LSTM ユニットによって学習する．各 LSTM ユニットからの出力ノードを結合層で結合し，中間層に入力する．LSTM ユニットが 2 層で，結合層と中間層が Linear の活性化関数をもつ．

Using Numeric and Textual Data for Dry Bulk Index Forecasting
†Ling Teng　†Tomoharu Nagao
‡Graduate School of Environment and Information Science, Yokohama National University

3. バルチック海運指数の予測実験

5日後と10日後のバルチック海運指数の予測実験を行い，提案手法の有効性を示す．

3.1 データセット

数値情報のデータは2014年から2016年までの4種類の海運指数を対象とする．それぞれ日次データで，バルチック海運指数（BDI），バルチック・ケープサイズ指数（BCI），バルチック・パナマックス指数（BPI），バルチック・スープラマックス指数（BSI）の海運指数値を使用する．週末以外の祝日に関しては，直近の指数値で補完処理を行う．過去 15 日間の各海運指数の指数値を数値特徴量とする．

テキスト情報のデータは2014年から2016年までの海運業界紙からの主要ニュース11397本を対象とする．それぞれのニュースのタイトルを形態素解析し，一日ごとにそれぞれの単語の頻度を求めた．2 つ単語の組合せが同時に出現する頻度とバルチック海運指数の変動率について，相関係数を算出した．相関係数の値が 0.3 以上ならば上昇関係の予測複合語とし，相関係数の値が-0.3 以下ならば下落関係の予測複合語とした．GA による探索結果から，5 日後の場合，3 つの予測複合語が抽出され，10 日後の場合，5 つの予測複合語が抽出された．これら過去 7 日間の特徴ベクトルをテキスト特徴量とする．

5日後の予測では，2014年4月から2016年5月までのデータ538件を学習し，2016年5月からの100件のデータで検証を行った．10日後の予測では，2014年5月から2016年5月までのデータ533件を学習し，2016年5月からの100件のデータで検証を行った．

3.2 実験結果

予測精度の評価として MAPE（Mean absolute percentage error）と方向一致率を用いた．比較手法として，数値情報だけを考慮する手法1とBoW（Bag of Words）によってテキスト情報のベクトル化を行う手法 2 を用いた．実験結果を表 1，図 2 に示す．

3.3 考察

表1の結果から，提案手法と手法2ではMAPEと方向一致率は両方とも数値のみを考慮する手法 1 より優れた結果を示した．数値情報とテキスト情報を用いることで予測精度の向上につながることが分かった．また，予測複合語の抽出を行う提案手法は BoW でテキストを表現する手法 2 よりも優れた精度を示した．相関性を考慮することで，有効な特徴量が抽出できたためであるからと考えられる．

表 1　予測結果の MAPE と方向一致率

	MAPE		方向一致率	
	5DAYS	10DAYS	5DAYS	10DAYS
手法 1	6.0%	9.0%	54%	51%
手法 2	4.9%	7.5%	54%	64%
提案手法	**4.4%**	**7.2%**	**64%**	**65%**

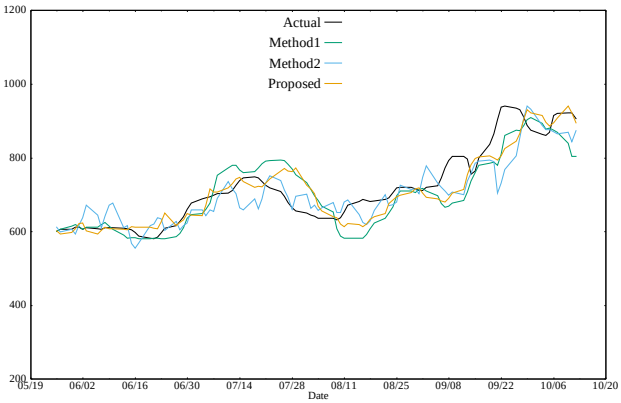

図 2　5 日後のバルチック海運指数の予測結果

4. まとめ

数値情報とテキスト情報を組み合わせたバルチック海運指数の予測手法を提案した．実際のバルチック海運指数の5日後，10日後の予測実験によって，提案手法はわずかであるが精度を向上できることが確認できた．今後は，より高精度な予測モデルの構築のために，特徴量を増やすことなどを行う予定である．

参考文献

[1] MG Kavussanos, AH Alizadeh: “Seasonality patterns in dry bulk shipping spot and time charter freight rates”. Transportation Research Part E 37(6) (2001),443–467.

[2] Han Q,Yan B,Ning G,Yu B: “Forecasting Dry Bulk Freight Index withImproved SVM”. Mathematical Problems in Engineering, (2014).

[3] Sepp Hochreiter, Jürgen Schmidhuber: “Long short-term memory”. Neural Computation 9 (8) (1997) 1735–1780.

[4] FA Gers, J Schmidhuber,and F Cummins: “Learning to forget: continualprediction with LSTM”. Neural Computation 12(10) (2000), 2451–2471.