2D1-5in1

時系列データの動向概要を示す要約文生成に向けて

Toward Automatic Summary Text Generation for the Trends of Time-series Data

青木　花純 *1　　小林　一郎 *2
Aoki Kasumi　　Kobayashi Ichiro

お茶の水女子大学
Ochanomizu University

Upon text generation from time-series data, the process of reducing the set of data to the language model has played a significant roll in increasing the effectiveness of our proposed method in our past studies. This paper describes and compares two different methods that can be applied to identify which linguistic resources will be given to the language model. For the time-series data, we use the price of Nikkei stock average and the program aims to generate linguistic texts which describes how the intraday stock price evolves in the market.

1.　はじめに

近年，ウェアラブル端末などのセンサ類の発達により容易に観測可能になった時系列数値データは，IOT への関心の高まりにより，様々な用途で利用する場面が増えている．しかし，時系列データをそのまま表示するだけでは人にとって理解が難しく，人の理解を助けるためにテキスト表現等を用いた動向概要を付与することが多く行われている [1, 2, 3, 7]．そのため，時系列数値データから動向概要を示すテキスト等を自動生成する技術への関心が高まっている．また，自然言語処理の分野においても，視覚情報として観測されるデータを時系列数値データとして処理し，テキスト生成する手法が盛んに研究されている．

本研究では，時系列数値データの動向概要を示す要約テキストを n-gram 言語モデルの資源を用いて生成すること目的としている．具体的に日経平均株価を例として，分類とクラスタリングの 2 つの手法を用いて言語モデル構築の際に使用する言語資源の特定手法について考察する．

2.　時系列データからの要約文生成

2.1　概要

本研究は，過去に観測された時系列数値データのパターンとその動向概要を示したテキスト内容の対応関係を学習することによって，観測された時系列数値データの動向概要を示すテキストを自動生成する言語モデルの構築を目的とし，言語モデルを構築する言語資源の特定のために提案した 2 つの手法を比較し，考察する．図 1 にテキスト生成の概要を示す．

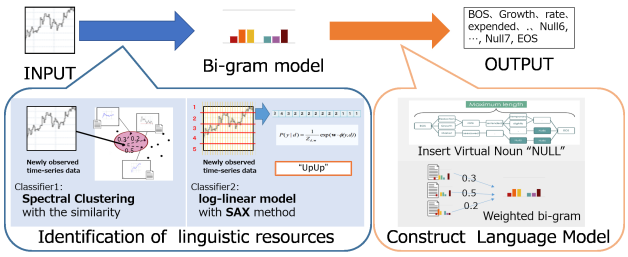


図 1: 提案手法概要

1 つ目の手法では，過去に観測された時系列データを学習データとして，新たに時系列数値データが与えられた際に尤もらしいテキスト内容 (以下，中間表現と呼ぶ) を決定する識別モデルを構築する．識別モデルの構築には対数線形モデルを用い，次に識別器によって識別された中間表現と同じ中間表現がつけられている過去の時系列データと対になっている株価の動向を説明する文書を言語資源として特定し，バイグラムモデルを構築する．

2 つ目の手法では，新たに観測された時系列数値データと過去に観測された時系列数値データに対してスペクトラルクラスタリングを適用し，任意の個数のクラスタに分類する．そして，新しく観測された時系列数値データと同クラスタに分類された各時系列数値データの動向内容を示した文書を言語資源として特定し，バイグラムモデルを構築する．その際，新しく観測された時系列数値データと同クラスタに分類された時系列データの類似度に応じて重み付けを行う．

以上の手法によって生成したバイグラムモデルに対し，動的計画法を用いて確率的に尤もらしい単語の組み合わせを決定し，観測された時系列数値データの動向概要を示すテキストを生成する．

2.2　識別器を用いた言語資源の特定

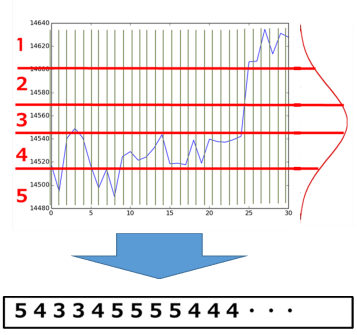


図 2: Symbolic Aggregation approXimation (SAX)

まず，図 2 のように過去に観測された時系列データを Symbolic Aggregation approXimation(SAX 法)[4] を用いて次元圧縮さ

れたデータに変換する．その後，圧縮された時系列データそれぞれに対して，人手で表 1 に示すような中間表現を付与する．

表 1: 中間表現とその意味内容

中間表現	意味内容
UU(UpUp)	上昇し続けた
DD(DownDown)	下落し続けた
SS(StayStay)	あまり変化はなかった
UD(UpDown)	上昇した後，下落した
DU(DownUp)	下落した後，上昇した
DS(DownStay)	下落した後，あまり変化しなくなった
SD(StayDown)	あまり変化していなかったが，その後下落した
US(UpStay)	上昇した後，あまり変化しなくなった
SU(StayUp)	あまり変化していなかったが，その後上昇した

そして新たなデータ d に対してその「中間表現」r を判定する識別モデルを式 (1) に示す対数線形モデルを用いて構築する．

$$P(r|d) = \frac{1}{Z_{d,\boldsymbol{w}}} exp(\boldsymbol{w} \cdot \phi(d, r)) \quad (1)$$

素性ベクトル ϕ は前場，後場における 5 分足の圧縮されたデータで構築されている．また，$Z_{d,\boldsymbol{w}}$ は正規化係数である．構築された識別器を用いて判定された中間表現 r_1 と同じ中間表現が付与されている過去の時系列データと対に収集した言語資源として特定する．この手法は人手でデータに中間表現を付与するためコストはかかるが，各中間表現の元で言語資源が管理されている．そのため，安定して言語資源の特定ができる．しあし，識別された中間表現に紐付いた言語資源のみを使用してテキストが生成されるため，テキストの表現には大きな多様性が見られない傾向にある．

2.3 クラスタリングによる言語資源の特定

スペクトラルクラスタリングはまず図 3 のように各データをノード，各データ間の類似度をノード間のエッジとして構築された無向類似度グラフをグラフカットしていくことでクラスタリングを行う手法である．

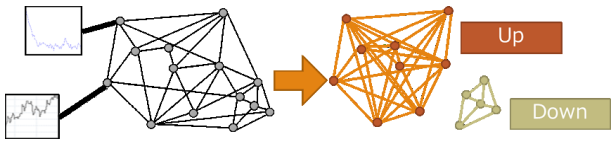


図 3: スペクトラルクラスタリング

本研究では，式 (2) に示される Normalized cut(NC) 関数を用いてグラフカットを行った．NC 関数はクラスタリングされる各クラスタ内と外のつながりを考慮した関数である．

$$Ncut(C_1, C_2) = \frac{cut(C_1, C_2)}{assoc(C_1, V)} + \frac{cut(C_1, C_2)}{assoc(C_2, V)} \quad (2)$$

ここで，

$$\text{assoc}(\text{C}_1, V) = \Sigma_{u \in C_1, t \in V} w(u, t),$$
$$\text{assoc}(\text{C}_2, V) = \Sigma_{u \in C_2, t \in V} w(u, t)$$

本研究では，時系列データ同士の類似度には DTW 距離と ψ-DTW 距離 [8, 9] を用いる．Algoritgm1 にスペクトラルクラスタリングのアルゴリズムを示す．時系列データの値域による類似度への影響を防ぐため，本研究では各時系列データを変化量に変換して利用する．

Algorithm 1 Spectral clustering algorithm

Input: $p = p_1, p_2, ..., p_n$:asset of time series data;k:number of clusters;$sim(p_i, p_j)$: similarity between p_i and p_j
Output: $C = C_1, C_2, ..., C_k$: Cluster
Initialize $W \in R^2$, $D \in R^2$
create similarity matrix W.
for all i such that $1 \leq i \leq n$ **do**
  for all j such that $1 \leq j \leq n$ **do**
    $W(i, j) = sim(i, j)$
  end for
end for
create diagonal matrix D by W.
for all i such that $1 \leq i \leq n$ **do**
  $D(i, i) = W(i, i)$
end for
compute normalized similarity matrix D.
and $S = D^{-1/2} W D^{-1/2}$
create k eigenvectors of S.
compute normalized matrix Y_k by using X_k.
do k-means clustering for n vectors of Y_k.
for all i such that $1 \leq i \leq n$ **do**
  add P_i to C_l.
end for
return $C = C_1, C_2, ..., C_k$.

2.3.1 時系列データの類似度

本研究では，時系列データの類似度として各時系列データの Dynamic Time Warping(DTW) 距離 [8] および Prefix and Suffix-Invariance DTW(ψ -DTW) 距離 [9] を用いた．

- **DTW 距離**
  DTW 距離とは，時系列データの各点の距離を総当りで計算し，距離行列を作成し，その距離行列を元に求めた時系列データ同士の距離が最短となるパスにおける総コストである．時系列データにおける挙動の周期性や挙動の長さが違う場合でも高い類似度を示す距離尺度として用いられている．window 幅などの制約がある．Algorithm2 に DTW 距離を求めるアルゴリズムを示す．

Algorithm 2 DTW:The standard DTW algorithm

Input: S: Sequence of length n, Q: Sequence of length m.
Output: DTW distance.
Initialize $C(i, j) = |S_i - Q_j|$
Initialize $D(i, 1) \Leftarrow i\delta$ for each i
Initialize $D(1, j) \Leftarrow j\delta$ for each j
for all i such that $2 \leq i \leq n$ **do**
  for all j such that $2 \leq j \leq m$ **do**
    $D(i, j) = C(i, j) + min(D(i-1, j), D(i-1, j-1), (i, j-1))$
  end for
end for
return $D(n, m)$

- **ψ-DTW 距離**
  従来の DTW 距離では実世界の時系列データにおいて適切な類似度を示せない場合がある．それは各時系列の区切りが適切でないことによるものである．その問題を解決するために提案された ψ-DTW は制約パラメータ r を DTW の距離に追加することによって，各時系列データの「Prefix/Suffix」の情報に左右されない類似度指標となっている．Algorithm3 に DTW 距離を求めるアルゴリズムを示す．

Algorithm 3 ψ-DTW

Input: x,y: sequences, relaxation factor parameter r
Output: ψ-DTW distance.
$M \Leftarrow$ infinity matrix $(n+1, m+1)$
Initialize $M([0,r],j)$ and $M(0,[0,r])$
for all i such that $1 \leq i \leq n$ **do**
  for all j such that $1 \leq j \leq m$ **do**
    $M(i,j) \Leftarrow \mathrm{c}(x_i, y_i) + min(M(i-1, j-1), M(i, j-1), M(i-1, j))$
  end for
end for
$minX \Leftarrow$min(M([n-r,n],m)),minY$\Leftarrow$min(M(n,[m-r,m]))
return $min(minX, minY)$

2.3.2 特定した言語資源の重み付け

言語モデルとして，観測された時系列データと同クラスタの言語資源を用いてバイグラムモデルを構築する．その際，観測された時系列データと同クラスタ内の各時系列データの類似度を基に各言語資源に重み付けを行い，バイグラムモデルを構築する．

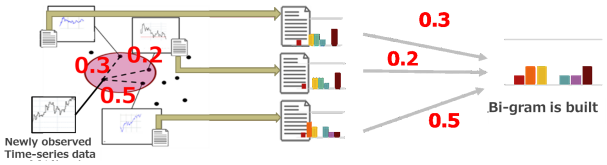
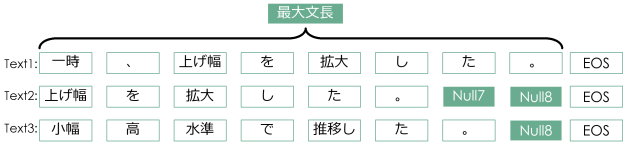


図 4: 時系列データの類似度に基づく言語モデル構築

2.4 言語モデルの構築および生成

以上の手法を用いて各言語モデルを構築し，動的計画法を用いることによって尤度が高くなる単語の組み合わせを獲得し，新たに観測された時系列データの動向概要を示す要約文を生成する．尤度は文長が長い文ほど低くなってしまうことから，文長に左右されない文生成を実現するため，言語モデルを構築する際には，図 5 のようにバイグラムモデルを構築する言語資源の最大文長に合わせて，各言語資源すべてに仮想の単語として番号付きの null ラベルを擬似単語として導入した．

図 5: 仮想単語 null の挿入



3. 実験

本章では，以上で提案した手法を用いて，新たに時系列数値データが与えられた際，その内容を説明する要約文を生成する実験を行い，評価を行う．

3.1 実験設定

対数線形モデルにおける識別ラベル数は中間表現の数に合わせて 3 もしくは 9 とし，スペクトラルクラスタリングにおけるクラスタ数は 3 18 と設定した．また株価の時系列数値データ，および言語モデルを構築する文書は前場，後場の各時間帯にわけて収集した．実験に使用したテキストデータ [*1]，および数値データ [*2] は，2013 年 2 月 25 日〜2014 年 12 月 30 日に収集された 451 日分の 902 個のデータを用いた．今回は収集したデータのうち，ランダムで選択したデータを新たに観測されたデータだとみなし，提案手法を適用した．その後，動的計画法を用いることで，株価数値データの概要を説明する尤もらしいテキストを生成した．

3.2 実行結果および考察

実行結果の例として，提案した 2 つの手法を用いて生成されたテキストを時系列数値データおよび正解文，そして各言語モデルのバイグラムのヒストグラムとともに表 2 に示す．

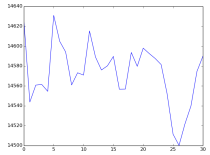
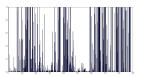
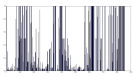
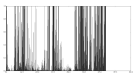
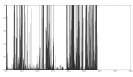
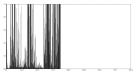
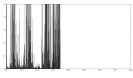
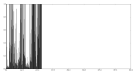
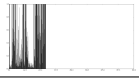
提案した 2 つの手法によって生成されたテキストには短文が多くみられた．その理由として，使用した過去のテキストデータに短文が多くみられたことが挙げられる．また生成された文には，非文がほとんど見られなかったが，意味的に冗長なテキストが生成されてしまうことがあった．そのため，過去に観測されたテキストのバイグラムのみによって言語モデルを構築するだけではなく，形容詞や動詞などの重要単語の組み合わせと文法を組み合わせる手法を用いた言語モデルの構築なども必要であると考えられる．生成された文の例のように，識別器を用いる手法によって生成されたテキストよりもクラスタリングを用いる手法によって生成されたテキストの方が自由度が高い文章が生成されている．一方で，SPC を用いる方法ではクラスタリングを予め設定する必要があり，設定したクラスタ数は言語資源の特定に大きな影響を与えると考えられる．そのため，今後はクラスタ数も推定するノンパラメトリックなクラスタリング手法が必要であると考える．

4. おわりに

本研究では，日経平均株価を例として，観測された時系列データの概要を説明するテキストの自動生成に取り組んだ．新たに与えられた時系列数値データに基づいて選択された言語資源を用いて，言語モデルを構築し，動的計画法を用いて尤度の高い単語の組み合わせを得ることでテキスト生成を行い，言語資源の特定には識別器を用いる手法とクラスタリングを用いる方法を用いた．今後の課題として，生成されたテキストの人手による評価や，クラスタリングを用いる手法においてクラスタ数を自動推定する手法の適用などが挙げられる．

*1 ADVFN:http://jp.advfn.com/より取得.
*2 IBI-Square Stocks:http://www.ibi-square.jp/より取得.

表 2: 生成されたテキストの一例

時系列データ		正解文		
		一時上げ幅を拡大した		
手法	クラスタ数	ラベル/アルゴリズム	正解文	bi-gram
対数線形モデル	3	識別結果：「U」	上げ幅, を, 拡大, し, た,。, ⋯,EOS	
	9	識別結果：「UU」	上げ幅, を, 拡大, し, た,。, ⋯,EOS	
SPC	3	DTW	上げ幅, を, 拡大, し, た,。, ⋯,EOS	
		ψ-DTW	上げ幅, を, 拡大, し, た,。, ⋯,EOS	
	9	DTW	一時, 上げ幅, を, 拡大, し, た,。, ⋯,EOS	
		ψ-DTW	一時, 下げ, 幅, を, 拡大, し, た,。, ⋯,EOS	
	18	DTW	一時, 上げ幅, を, 拡大, し, た,。, ⋯,EOS	
		ψ-DTW	一時, 下げ, 幅, を, 拡大, し, た,。, ⋯,EOS	

参考文献

[1] Gkatzia, D., Hastie, H. and Lemon, O., Finding middle ground Multi-objectuve Natural Language Generation from time-series data, the 14th European Association for Computational Linguistics, pp.210-214,2014.

[2] H., Banaee, M. U. Ahmed, A. Loutfi, A Framework for Automatic Text Generation of Trends in Physiological Time Series Data, IEEE Int. Conf. on Systems, Man, and Cybernetics, pp.3876-3881,2013.

[3] 小林瑞希, 小林一郎, 麻生英樹, 同画像中の人の動作を表現する確率的言語生成に関する取り組み (2013). 第 27 回人工知能学会全国大会,2D5-OS-03b-3, 2013.

[4] J.Lin, E. Keogh, S. Lonardi, Experiencing SAX: a Novel Symbolic Representation of Time Series., Springer Science+ Business Media, LLC 2007.

[5] Ulrike von Luxburg ”A Tutorial on Spectral Clustering” Max Planck Institute for Biological Cybernetics Spr,spemannstr. 38, 72076 Tubinge, Germaniy, Statics and Computing 17 (4),2007.

[6] Inderjit Dhillon, Yuqiang Guan, and Brian Kulis, A Unified View of Kernel k-means, Spectral Clustering and Graph Cuts, In The University of Texas at Austin, Department of Computer Science. Technical Report TR-04-25,2005.

[7] 青木花純, 小林一郎, 時系列データのパターンを考慮した言語モデルに基づく自然言語生成, 情報処理学会,2016.

[8] E. J. Keogh. Exact indexing of dynamic time warping. In Proceedings of VLDB, pp. 406417, Hong Kong, China,August 2002.

[9] Silva, D. F., Batista G. E. A. P. A. and Eamonn Keogh, “ Prefix and Suffix Invariant Dynamic Time Warping ”, in Proceedings of the IEEE International Conference on Data Mining, 2016.