1F5-2in1

クラスタリングを用いた時系列数値データの動向内容を示すテキストの自動生成手法への一考察

A Study on Automatic Text Generation for Describing the Trends of Time-series Data using spectral Clustering

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

お茶の水女子大学
Ochanomizu University

This paper describes a method to verbalize the trends of time-series data. As an example of time-series data, we use the price of Nikkei stock average and develop a method to generate texts which describe how the stock price goes in the market. As the basic idea for making linguistic descriptions of the stock price trends, we firstly classify all the time-series data including a newly observed time-series data, i.e., the target to be verbalized, by means of spectral clustering employing Dynamic Time Warping distance as its similarity metric. Secondly, a bi-gram language model for the newly observed data is built by weighting the bi-gram language models of the other time-series data classified in the same cluster according to he similarity to the target data. Lastly, linguistic summarization for the target data is generated by finding the most likely combination of words be means of dynamic programming, employing the weighted bi-gram model.

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時系列数値データが容易に観測可能になった．これらの時系列数値データを様々な用途で利用する場面が増えているが，時系列データをそのまま表示するだけでは理解が困難であり，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．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 [1, 2, 3, 6]．本研究では，特にグラフとして表現される時系列データを取り上げ，日経平均株価を例に，時系列データの類似度を基に重み付けされた言語モデルを生成し，その言語モデルに基づき，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手法を提案する．

2.　クラスタリングに基づくテキスト生成

本研究では，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文章内容の言語資源を利用することによって，観測された時系列数値データの動向概要を表現するテキストを自動生成する．図 1 に提案手法の概要を示す.

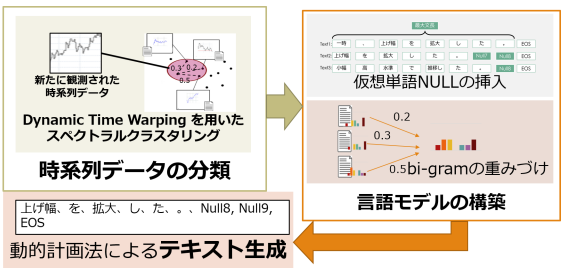


図 1: 提案手法概要

まず，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 Dynamic Time Warping(DTW) 距離 [7] を時系列データ同士の類似度としてスペクトラルクラスタリング [4, 5] 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からバイグラムモデルを構築する．その際に，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.
以上のように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.

2.1　時系列データの分類

スペクトラルクラスタリングは各データをノード，各データ間の類似度をノード間の距離として Normalized Cut を行う事で，各データをクラスタリングする手法である．本研究では，時系列データ同士の類似度には DTW 距離を用いた. DTW 距離とは時系列データの各点の距離を総当りで比較し，総計コストが最短となるパスでかかる総コストのことである. 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した.

Algorithm 1 DTW:The standard DTW algorithm

Input: S: Sequence of length n, Q: Sequence of length m.
Output: DTW distance.
$InitializeD(i,1) \Leftarrow i\delta$ for each i
$InitializeD(1,j) \Leftarrow i\delta$ for each i
for all i such that $2 \leq i \leq n$ **do**
  for all j such that $2 \leq j \leq n$ **do**

$$\mathrm{D(i,j)} = \mathrm{d(i,j)} + \min \begin{cases} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{cases}$$

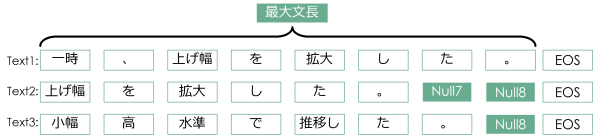
  end for
end for
return $D(n,m)$

2.2　言語モデルの構築および生成

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に分類された時系列数値データの動向概要を示すテキストを言

語資源として用い, 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（DTW 距離）を基に各言語資源に重み付けを行い，バイグラムを構築した. このように構築したバイグラムに動的計画法を用いて，尤度が高くなる単語の組み合わせを得ることで, 確率的に尤もらしい文を生成する．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テキスト生成を実現するため，図 2 のように仮想の単語 null を挿入した．また計算された尤度に対して，単語の perplexity を加味した平均情報量を計算し，情報量の大きい文を生成文として出力した.

図 2: 仮想単語 null の挿入



3. 実験

本章では，上記に説明した手法を用いて，新たな日経平均株価の時系列数値データが与えられた際，その内容を説明するテキスト生成の実験を行い，評価を行う.

3.1 実験設定

今回使用する日経平均株価は動向内容が上昇，下落後安定など，おおよそ 9 個に経験的に分類できると仮定し，実験ではその数の前後の数に分類されると想定し，時系列数値データは 6〜12 個にクラスタリングされるとした.

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章は前場，後場の各時間帯に分けて収集した. 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 個の前場，後場のデータである．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 (評価用データに相当) とみなし，提案手法を適用した.

3.2 実験結果

提案手法を用いて構築したバイグラムに動的計画法を用いることで数値データの動向概要を示す尤もらしい文を生成した. クラスタリングを行った分類結果および，言語資源の統計値を表 1 および表 2 にそれぞれ示す．また生成された文の例を表 3 に，生成文に対する評価値を 4 に示す.

表 1: 時系列データの分類

クラスタ数												
6	100	113	77	51	74	73	–	–	–	–	–	–
7	72	77	97	57	29	88	68	–	–	–	–	–
8	59	41	66	89	55	83	50	45	–	–	–	–
9	63	56	51	52	50	57	55	52	52	–	–	–
10	58	30	59	49	66	48	46	44	40	48	–	–
11	33	43	49	35	74	31	37	61	43	37	45	–
12	37	42	37	49	49	38	40	24	26	57	52	37

3.3 考察

クラスタ数によらず，DTW 距離の平均値が高いことから，クラスタ分割が機能していることがわかった．また，クラスタ数を変化させることで生成文には違いが見られ，クラスタ数が大きいものほど，適切な表現ができているように感じた．しかし，評価値には大きな差異が見られなかった．これは使用した評価指標では単語の含意関係や，文中の単語の前後関係の考慮がされていないこと，そして動向説明文の文長が固定されていることが原因であると考えられる.

∗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: 言語モデルの統計量

評価値/クラスタ数	6	7	8	9	10	11	12
単語の種類数	147	145	133	151	116	115	113
バイグラムの種類数	344	334	285	315	230	237	230
DTW 平均値	0.66	0.62	0.68	0.61	0.70	0.60	0.58

表 4: 生成文の評価値

評価値/クラスタ数	6	7	8	9	10	11	12
precision	0.53	0.54	0.52	0.52	0.52	0.52	0.53
recall	0.37	0.37	0.36	0.36	0.36	0.37	0.37
f1 value	0.41	0.42	0.41	0.40	0.40	0.41	0.42

4. まとめ

本研究では，日経平均株価を対象に，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．時系列数値データに対して類似度に基づくクラスタリングを行い，同クラスタに分類された時系列データのバイグラムに類似度を重み付けし，言語資源としてバイグラムを構築し，そのバイグラムに対して動的計画法を用いることにより，尤度の高い単語の組み合わせを得ることで文生成を行った．時系列数値データの分類におけるクラスタ数や言語モデルを構築の際の重み付け方法を比較し，考察を行った．クラスタ数が多いものほど適切な文が生成できていたが，評価値にはあまり差異は見られなかった．今後はクラスタ分割数の決定や言語資源への重み付けなどを改善し，精度の高い文を生成したいと考えている．また，生成文の評価方法を検討したいと考えている.

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.

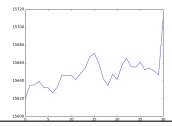
[4] Ulrike von Luxburg ”A Tutorial on Spectral Clustering” Max Planck Institute for Biological Cybernetics Spr,spemannstr. 38, 72076 Tubinge, Germaniy, Statics and Computing 17 (4),2007.

[5] 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.

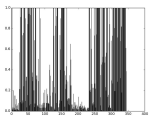
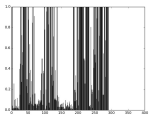
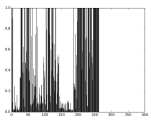
[6] 青木花純, 小林一郎, 時系列データのパターンを考慮した言語モデルに基づく自然言語生成, 情報処理学会,2016.

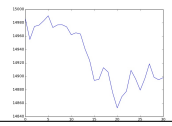
[7] E. J. Keogh. Exact indexing of dynamic time warping. In Proceedings of VLDB, pp. 406417, Hong Kong, China,August 2002.

表 3: 言語モデルによる生成文例

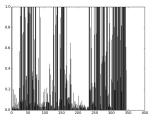
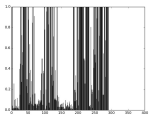
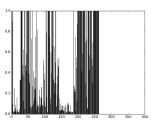


正解文
一進一退の相場となった。

クラスタ数	生成文	対数尤度	バイグラム
6	上げ幅, を, 拡大, し, た, 。, null8, ···, null36, EOS	-152.01	
	一時, 上げ幅, を, 拡大, し, た, 。, null9, ···, null36, EOS	-154.22	
	一時, 下げ, 幅, , 拡大, し, た, 。 , null10, ···, null36, EOS	-156.01	
9	上げ幅, を, 拡大, し, た, 。, null8, ···, null27, EOS	-120.22	
	一時, 上げ幅, を, 拡大, し, た, 。, null9, ···, null27, EOS	-122.40	
	一時, 下げ, 幅, を, 拡大, し, た, 。, null10, ···, null27, EOS	-123.88	
12	上げ幅, を, 拡大, し, た, が, ,, 下げ, 幅, を, 拡大, し, た, 。, null16, ···, null25, EOS	-112.63	
	下げ, 幅, を, 拡大, し, た, が, ,, 下げ, 幅, を, 拡大, し, た, 。, null17, ···, null25, EOS	-113.73	
	下げ, 幅, を, 拡大, し, た, が, ,, 下げ, 幅, を, 拡大, し, た, 。, null16, ···, null25	-115.60	



正解文
下げ幅を拡大した。

クラスタ数	生成文	対数尤度	バイグラム
6	上げ幅, を, 拡大, し, た, 。, null8, ···, null36, EOS	-133.90	
	一時, 上げ幅, を, 拡大, し, た, 。 , null9, ···, null28, EOS	-135.01	
	一時, 下げ, 幅, を, 拡大, し, た, 。 , null10, ···, null28, EOS	-137.07	
9	下げ, 幅, を, 拡大, し, た, 。, null9, ···, null29, EOS	-117.75	
	一時, 下げ, 幅, を, 拡大, し, た, 。, null10, ···, null29, EOS	-122.40	
	一時, 下げ, 幅, を, 拡大, し, た, 。, null9, ···, null29, EOS	-120.10	
12	一時, 下げ, 幅, を, 拡大, し, た, 。, null10, ···, null25, EOS	-121.93	
	一時, 下げ, 幅, を, 拡大, し, た 。, null9, ···, null25, EOS	-123.14	
	一時, 下げ, 幅, を, 拡大, し, た, 。, null8, ···, null25	-124.61	