4K1-OS-2-1

複数の時系列データの比較に基づく言語化の試み

An Approach to Linguistic Summarization based on Comparison among Multiple Time Series Data

小林 瑞季*1　　小林 一郎*1
Mizuki Kobayashi　　Ichiro Kobayashi

*1お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻 情報科学コース
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

This paper proposes a method of linguistic summarization of the relation among multiple time-series data by comparing them. The relation among the data is found by correlation coefficient and then it is categorized into main three relations: (i) the same trends, (ii) synmmetrical trends, and (iii) no correlation. Symbolic Aggregate approXimation (SAX) is applied to the data categorized into these three types for coding numerical data, and then significant parts between objective two time-series data are extracted by the extended edit distance we have proposed.

1. はじめに

気温や気圧，心電図や脳波，株価や為替など私たちの身の回りで観測されるデータの多くは時系列データである．そういった時系列データの振る舞いをより分かりやすく伝えるため，データを可視化によって表現することは多く見られる．しかし，大量の時系列データを扱う際，個々の時系列データの振る舞いだけでなく，それぞれの時系列間の関連性をみる必要性が生じ，複数の時系列データの関係を視覚的に俯瞰するのは難しくなる．

そのような背景から，本研究では，複数の時系列データを比較し，それらの関係をわかりやすく言葉で説明することを目的とする．具体的なアプローチとして二つの時系列データの相関係数をとることより，おおまかに，(i) 類似の動きをするもの，(ii) 対称の動きをするもの，(iii) 関連性がないもの，の 3 つのタイプに分類する． それぞれの分類に対して，SAX 法 [5] を用いて数値データを記号化し，編集距離 [7] を拡張し，2 つのデータ間で特徴的な箇所を抽出し，言語で表現する．

2. 関連研究

時系列データの言語化への取り組みとして，末吉ら [1][2] は，時系列データが状況や文脈によって柔軟に解釈されるために，データに対して最大値や最小値，上昇・下降・安定といった特徴を付与し，特徴の重要度に応じて付けられた重みを変更することにより異なるデータや期間でも柔軟にデータの特定部を比較できる枠組みを提案している．関ら [3] は，日経平均株価の時系列データを対象に，データの動向を巨視的に捉え言語化する手法を提案している．言語化においては，対象とするデータを説明するコーパスを分析し，時系列データの振る舞いを説明する語彙を抽出し，それらの語彙を用いて観測されたデータの特徴を言語化している．

また，時系列データの解析においては，金城 [4] らは，局所モデリングを行う時系列データマイニングにより，類似した時系列モデルの発見に基づき，意味のあるパタンを抽出する手法を提案している．大西ら [8] は，センサによって観測された時系列データからイベント検出のために SAX 法を用いたデータのインデックス化を行っている．

これらに対し，本研究では，まず複数の時系列データに対し相関関係を調べ，比較すべき時系列データを発見し，その後，改良した編集距離に基づいた SAX 法により，時系列データの動向を捉え特徴ある箇所を絞り込んだ後に言語化する手法を提案している．

連絡先: 小林瑞季，お茶の水女子大学大学院人間文化創成科学研究科理学専攻情報科学コース，〒112-8610 東京都文京区大塚 2-1-1，03-5978-5708，kobayashi.mizuki@is.ocha.ac.jp

3. 時系列データ間の特徴点抽出

3.1 SAX 法

SAX(Symbolic Aggregate approXimation)[5] とは，時系列データの近似表現方法の 1 つで，時系列データを文字列に変換する方法である．SAX を行う際，まず PAA(Piecewise Aggregate Approximation) というデータ圧縮作業を行う．長さ n の時系列データ C を用いて，w 次元の空間ベクトル $\bar{C} = \bar{c_1}, \ldots, \bar{c_w}$ に変換すると仮定する．$\bar{C}$ の i 番目の要素は式 (1) を用いて計算される．

$$\bar{C}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} C_j \qquad (1)$$

つまり，データを等間隔に w 個のフレームに分け，それぞれのフレーム内でのデータの平均をとることで，n 個ある時系列データを w 個の要素に簡約することができる．正規分布に従って，a,b,c…とアルファベットを割り振り，正規分布の各面積が等しくなるような分割線を定める．先ほど求めた平均値をこの分割線に従って文字に変換する (図 1 参照)．

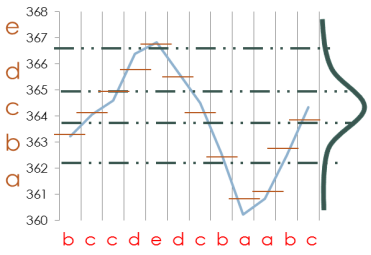


図 1: SAX 法による文字列変換

3.2 編集距離

SAX 法を用いて抽出された記号列は一般に，編集距離 (Levenshtein Distance) [7] という指標を用いて比較される．編集距離では 2 つの文字列 R，S に対し，置換・挿入・削除を行うことで文字列 R から文字列 S に変換する際にかかるコストの最小の合計値を文字列 R，S 間の距離とした，2 つの文字列間の類似度の指標として扱う手法である．

例として，2 つの文字列 “little”, “letter” に対して編集距離を求める．この場合，置換・削除・挿入の作業をそれぞれ一回ずつ行えば，“little” から “letter” に変換することができる (図 2 参照)．

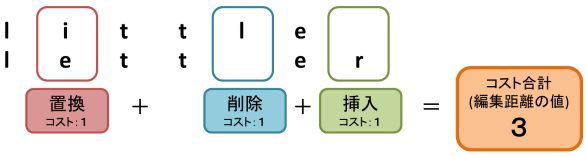


図 2: “little” と “letter” の編集距離のイメージ

編集距離を求める際には，動的計画法 (Dynamic Programming) が用いられる．文字数がそれぞれ N_R，N_S の場合は $(N_R+1) \times (N_S+1)$ 行列を用い，文字列 R の i 番目と文字列 S の j 番目 の編集距離 $LD(R_i, S_j)$ は 1 つ前までの編集距離 $LD(R_{i-1}, S_j), LD(R_i, S_{j-1}), LD(R_{i-1}, S_{j-1})$ によって式 (2) のように求められる．

$$LD(R_i, S_j) = \begin{cases} LD(R_{i-1}, S_j) + dist(r_{i-1}, s_j) & (\text{削除}) \\ LD(R_i, S_{j-1}) + dist(r_i, s_{j-1}) & (\text{挿入}) \\ LD(R_{i-1}, S_{j-1}) + dist(r_i, s_j) & (\text{置換}) \end{cases} \tag{2}$$

ここで $dist(r_x, s_y)$ とは，置換・挿入・削除それぞれにかかるコストである．ここでは，全てのコストを 1 として計算をする．

以上により，上記で例に出した “little” と “letter” の場合は以下のようになり，編集距離は 3 であることが分かる (図 3)．

		l	i	t	t	l	e
	0	1	2	3	4	5	6
l	1	0	1	2	3	4	5
e	2	1	1	2	3	4	4
t	3	2	2	1	2	3	4
t	4	3	3	2	1	2	3
e	5	4	4	3	2	2	2
r	6	5	5	4	3	3	3

図 3: “little ” と “letter” の編集距離

3.3 編集距離の拡張

通常の編集距離は，対応する個々の記号の比較において，記号が異なる数，または，記号を一致させるのに必要なコストを 2 つの時系列データの距離 (差異) としていたが，本研究では値そのものの比較よりも動きの比較の方にに重きを置いているため，上記の編集距離の手法をもとに，記号列の動向を比較し，同じ動向を持つ記号列に変更するのに要する置換・挿入・削除のコストを新たに編集距離として採用する (図 4 中，アルファベットの下の数値がそれぞれの時系列データの動向を示す)．

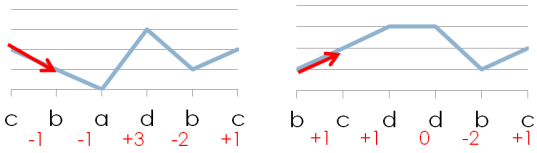


図 4: 編集距離の拡張

また，抽出されたこの 2 つの数値列をマッチングすることによって動向を比較し，2 つの時系列データ間の以下に示す 2 つの関係を取得する．

1. 類似した動き
-動きを示す値が全く同じ箇所．または，動きを示す値が，正 (上昇) なら「+」を，負 (下降) なら「−」を，0(一定) なら「0」を当てはめ，その記号が同じ箇所 (図 5 中，赤枠参照)．

2. 対称の動き
-正負は違うものの，動きを示す値の絶対値が全く同じ箇所．または，動きを示す値が，正 (上昇) なら「+」を，負 (下降) なら「−」を当てはめ，その記号が全く逆の箇所 (図 6 中，赤枠参照)．

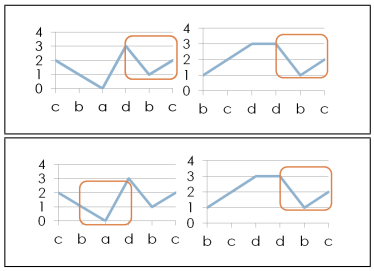


図 5: 類似した動き

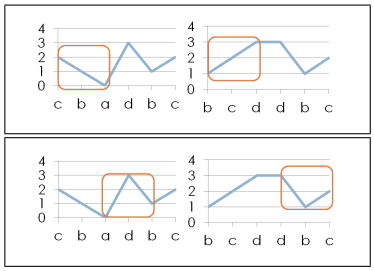


図 6: 対称の動き

3.4 相関関係に基づく特徴点抽出

2 組の数値からなるデータ列 $(x, y) = (x_i, y_i)(i = 1, 2, \ldots, n)$ が与えられたとき，相関係数は式 (3) で表される．

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3}$$

時系列データの相関関係は，相関係数の値により次の 3 つのタイプに分類される．

(i) 相関係数が正に高い組
(ii) 相関係数が負に高い組
(iii) 相関係数の低い組．

ここで，(i) は類似の動きをする時系列データ，(ii) は対称の動きをするデータ，(iii) は関連性が無いとされる時系列データ

であることを示す．本研究では (i)(ii)(iii) において，比較する時系列データから特異な特徴を抽出することにより，ユーザに時系列データの比較において新たな気付きを与えることを目指す．

上記 (i)(ii)(iii) に対する抽出方法をそれぞれ示す．

(i) 相関が正に高い組

おおよそ類似するデータ間において対称に類似する箇所の抽出を行う．

まず SAX によるフレームの間隔を大きくとり，編集距離を用いたマッチングをすることによって，大まかに見て「類似した動きをする箇所」を抽出する．その後フレームの間隔小さくとり，また編集距離を用いたマッチングをすることによって，部分的に「対称の動きをする箇所」を抽出する．

(ii) 相関が負に高い組

おおよそ対称に類似するデータ間における類似箇所の抽出を行う．

(i) 相関が正に高い組と同様にまず SAX によるフレームの間隔を大きくとり，編集距離を用いたマッチングをすることによって，大まかに見て「対称の動きをする箇所」を抽出する．その後 SAX のフレーム間隔を小さくとり，また編集距離を用いたマッチングをすることによって，部分的に「類似した動きをする箇所」を抽出する．

(iii) 相関が低い組

関連性の低いデータ間における類似箇所また対称に類似する箇所の抽出を行う．

まずデータを時間軸上に細かく分けそれぞれ相関係数を取り，部分的に相関の高い箇所を見つける．その箇所に対し SAX によるフレームの間隔を小さくとり，編集距離を用いたマッチングをすることことによって，部分的に「類似した動きをする箇所」また「対称の動きをする箇所」を抽出する．

4. 抽出された特徴点の言語化

3.4 節の抽出によって得られた特徴点を，3.3 節で定義した編集距離を用い，あらかじめ用意したテンプレートに当てはめ言語化する．以下にその例を示す (図 7 参照)．

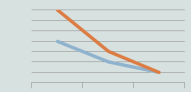
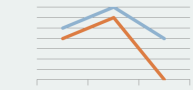
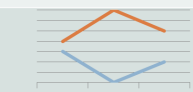
テンプレート例	条件	条件に合う編集距離の例
どちらも下落の動きを示し，[データ 1]の方が下げ幅が大きい	編集距離の値がどちらも全て負の値である下落の動きで「類似の動きをしている」と判断され，また[データ 1]の方が編集距離の値の絶対値が大きかった場合	 データ 1 [-4,-2], データ 2 [-2,-1]
どちらも同様に上昇ののち下落する動きを示しているが，[データ 1]の方が下落の下げ幅が大きい	編集距離の値がどちらも正ののち負の値をとる山形の動きで「類似の動きをしている」と判断され，また編集距離の負の値のみが異なり[データ 1]の方がその値の絶対値が大きかった場合	 データ 1 [2,-6], データ 2 [2,-3]
[データ 1]は上昇ののち下落する動きを示しているが，[データ 2]は逆に下落ののち上昇する動きを見せている	「対称の動きをしている」と判断され，[データ 1]の編集距離が正ののち負の値をとる山形の動きで，[データ 2]の編集距離が負ののち正の値をとる谷形の動きをする場合	 データ 1 [3,-2], データ 2 [-3,2]

図 7: 言語化に使用するテンプレートの例

5. 実験

以下に実験の内容をその手順に従って示す．

step 1. データ入力

複数の時系列データをデータベースに入れる．

ここでは，2011 年 12 月 5 日の日経平均 17 業種別株価 (全 18 個) について，それぞれ，9:00〜15:00 (休憩時間 11:30〜12:30) を 5 分足でとってきた時系列データ (データ数：62) を使用する (図 8 参照)．

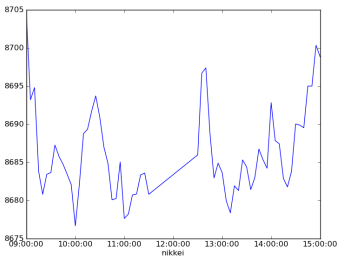


図 8: 使用するデータの一例

step 2. 相関係数によるタイプ分け

式 (3) で相関係数を求め，あらかじめ設定した閾値に基づき 3 つのタイプ，(i) 類似の動きをするもの，(ii) 対称の動きをするもの，(iii) 関連性がないもの，に分類する．

step 3. 比較

3.4 節で述べたように，それぞれのタイプに対し SAX 法と拡張した編集距離を用いてマッチングを行い，それぞれ「類似の動きをする箇所」「対称の動きをする箇所」を抽出する．

step 4. 言語化

発見されたデータ間の特徴点を言語化のためのテンプレートに照らし合わせて，文章とグラフを用いて表示する．図 9 にて，(iii) 相関係数の低いペアから「類似の動きをしている」と判断され，グラフと文章によって力された例を，また図 10 に，(iii) 相関係数の低いペアから「対称のの動きをしている」と判断され，グラフと文章によって出力された例を示す．

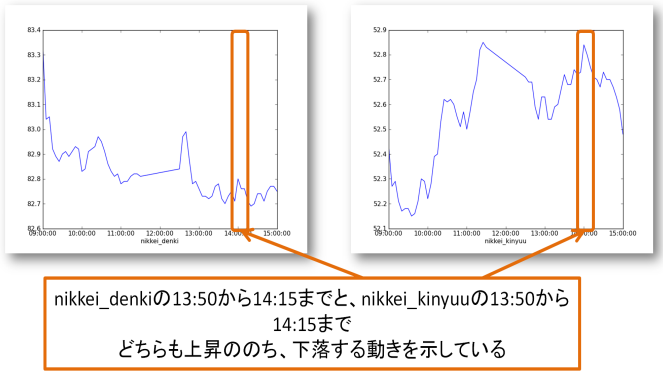


図 9: 相関関係の低いデータ間で類似した動きの言語化例

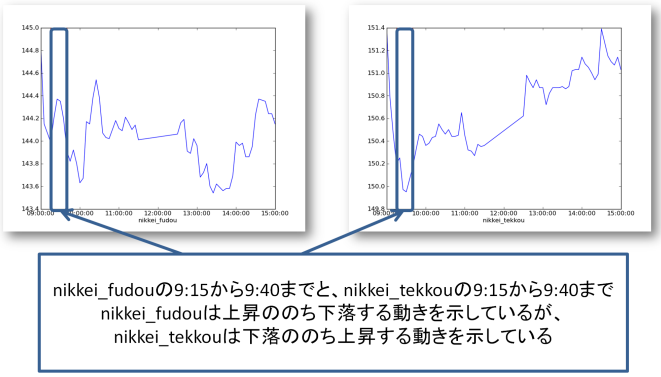


図 10: 相関関係の低いデータ間で対称の動きの言語化例

6. 被験者実験

5. 節で行った実験の結果について，出力された言語表現と実際のデータのふるまいが一致しているかについて被験者実験を行い確認した.「とても一致している」「一致している」「どちらともいえない」「一致していない」「まったく一致していない」の 5 択で 20 代女性 13 人を対象に，以下に示すアンケートを行った (図 11 参照).

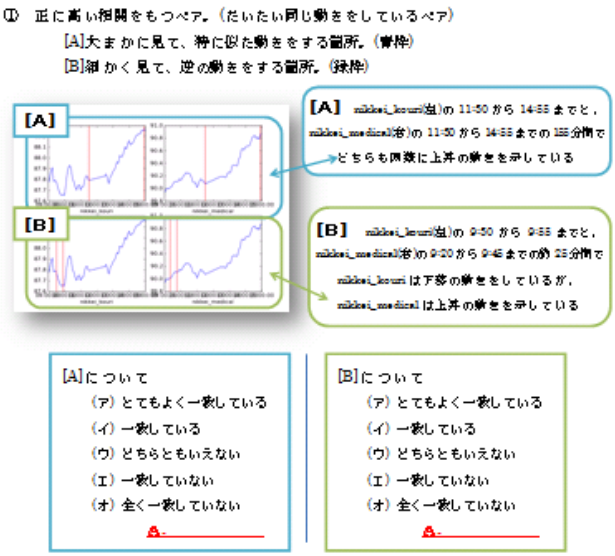


図 11: アンケート質問紙例

集計結果を以下に示す (図 12 参照).

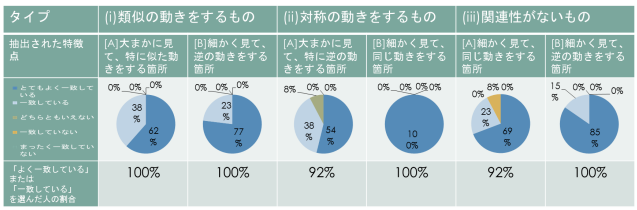


図 12: アンケート結果

どちらの項目も,「とても一致している」「一致している」と答えた方が 9 割を超えている．これにより，実際のデータのふるまいを言語によって表わすことができていると考えられる．

7. おわりに

本研究では，複数の時系列データを相関係数に基づき比較し，拡張した編集距離を用いた SAX 法を用いることにより，対象とする時系列データ間において特徴ある箇所を言語化する手法を提案した．

今後，より正確な結果を表示させるため分析方法を見直すとともに，パーティクルフィルタなどを取り入れて，データの予測を行い、その予測された動向についての関連性なども含め言語化をしていきたいと考える．また，あらゆる動きに柔軟に対応できるテンプレートの作成を目指すとともに，実際のニュース記事などを取り入れることにより言語表現の幅をさらに広げていきたいと考える．

参考文献

[1] 末吉れいら, 田中和広, 白水菜々重, 松下光範: 比較対象に着目したグラフの言語表現の生成, 第 21 回 Web インテリジェンスとインタラクション研究会, pp. 3738 (2011).

[2] 末吉れいら, 松下光範, 白水菜々重: 複数の時系列データの比較に基づくグラフの言語表現生成手法，人工知能学会 インタラクティブ情報アクセスと可視化マイニング研究会 (第 1 回) SIG-AM-01-03

[3] 関亜沙美，小林一郎：時系列データの言語化への取り組み－日経平均株価を例として－，第 24 回人工知能学会全国大会, 2010

[4] 金城敬太，澤井啓吾，古川康一：局所モデリング時系列データマイニングと帰納論理による知識獲得，第 20 回人工知能学会全国大会, 2007

[5] Lin, J. et al. Lin, J., Keogh, E., Lonardi, S. and Chiu, B. : A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, DMKD' 03, 2003.

[6] Levenshtein VI. “Binary codes capable of correcting deletions, insertions, and reversals ”Soviet Physics Doklady, 1996.

[7] Lei Chen and Raymond Ng:”On the marriage of L p-n orms and edit distance,” In Thirtieth International Conference on Very Large Data Bases (VLDB 2004), 2004.

[8] 大西 史花，渡辺 知恵美：スマートハウスのセンサデータに対する SAX を利用したイベント検出の検討，DEIM Forum 2011