複数の時系列データの関連性発見に基づく言語化の一考察

小林 瑞季　(指導教員：小林 一郎)

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 [1] を用いて数値データを記号化し，編集距離 [2] を拡張し，2 つのデータの特徴的な箇所を抽出し，言語で表現する．

2　時系列データ間の特徴点抽出

2.1　SAX 法

SAX(Symbolic Aggregate approXimation)[1] 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)．

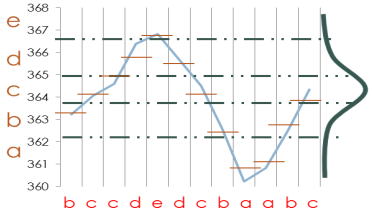


図 1: SAX 法による文字列変換

2.2　編集距離の拡張

SAX 法を用いて抽出された記号列は，編集距離 (Levenshtein Distance) [2] という指標を用いて比較される．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) とするが，本研究では，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コストを新たに編集距離として採用する (図 2 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)．

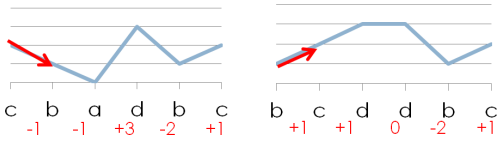


図 2: 編集距離の拡張

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 4 つの関係を取得する．

1. 同じ動き
　-動きを示す値が全く同じ箇所 (図 3 中，枠参照)．

2. 類似した動き
　-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 4 中，枠参照)．

3. 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 (図 5 中，枠参照)．

4. 対称に類似した動き
　-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，枠参照)．

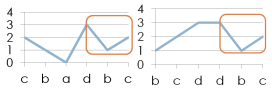


図 3: 同じ動き

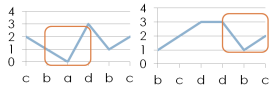


図 4: 類似した動き

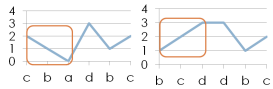


図 5: 対称の動き

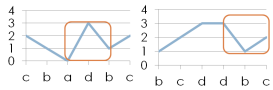


図 6: 対称に類似した動き

2.3　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 (2) で表される．

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

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ることによって，大まかに見て「同じ動きをする箇所」「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，部分的に「対称の動きをする箇所」「対称に類似した動きをする箇所」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) 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ることによって，大まかに見て「対称の動きをする箇所」「対称に類似した動きをする箇所」を抽出する．その後 SAX のフレーム感覚を小さくとり，部分的に「同じ動きをする箇所」「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ることによって，部分的に「同じ動きをする箇所」「類似した動きをする箇所」また「対称動きをする箇所」「対称に類似した動きをする箇所」を抽出する．

3 抽出された特徴点の言語化

2.3 節の抽出によって得られた特徴点を，2.2 節で定義した編集距離を用い，あらかじめ用意した 40 個の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)．

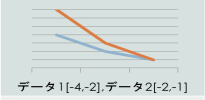
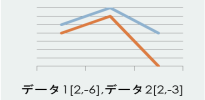
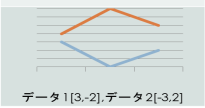
テンプレート例	条件	条件に合う編集距離の例
どちらも下落の動きを示し，[データ 1]の方が下げ幅が大きい	編集距離の値がどちらも全て負の値である下落の動きで「類似の動きをしている」と判断され，また[データ 1]の方が編集距離の値の絶対値が大きかった場合	 データ 1[-4,-2],データ 2[-2,-1]
どちらも同様に上昇ののち下落する動きを示しているが，[データ 1]の方が下落の下げ幅が大きい	編集距離の値がどちらも正ののち負の値をとる山形の動きで「類似の動きをしている」と判断され，また編集距離の負の値のみが異なり[データ 1]の方がその値の絶対値が大きかった場合	 データ 1[2,-6],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

4 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 17 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．

step 2. 相関係数によるタイプ分け
式 (2) で相関係数を求め，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
2.3 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，それぞれの特徴点を抽出する．

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 8 に例を示す．

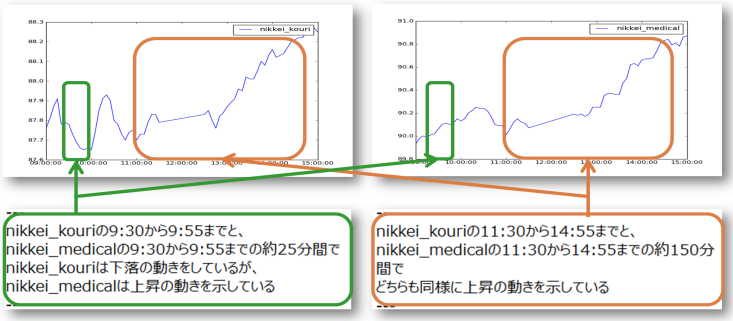


図 8: 特徴点を示す文章とグラフの対応

5 被験者実験

4 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，「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 13 人を対象にアンケートを行った (図 9 参照)．

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	
抽出された特徴点	[A]大まかに見て、特に似た動きをする箇所	[B]細かく見て、逆の動きをする箇所	[A]大まかに見て、特に逆の動きをする箇所	[B]細かく見て、同じ動きをする箇所	[A]細かく見て、同じ動きをする箇所	[B]細かく見て、逆の動きをする箇所
とてもよく一致している / 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	 0% 0% 0% 38% 62%	 0% 0% 0% 23% 77%	 8% 0% 0% 38% 54%	 0% 0% 0% 0% 10% 0%	 0% 8% 0% 23% 69%	 15% 0% 0% 0% 85%
「よく一致している」または「一致している」を選んだ人の割合	100%	100%	92%	100%	92%	100%

図 9: アンケート結果

6 おわりに

本研究では，複数の時系列データを比較することによりそれらの関連性を分かりやすく言葉で説明する手法を提案した．今後，意味を為さない特徴点の表示を減らすため分析方法を見直すとともに，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．

参考文献

[1] 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.

[2] Levenshtein VI. "Binary codes capable of correcting deletions, insertions, and reversals "Soviet Physics Doklady, 1996.