

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 小林一郎
お茶の水女子大学 理学部 情報科学科

1 はじめに

本研究では複数の時系列データを分析することによりその関連性を捉え、わかりやすく言葉で説明することを目的とする。時系列データの分析には SAX 法 [1] を用いて数値データを記号化し、その記号の比較に基づき時系列データ間の関連性を捉える。個々の時系列データから抽出された記号列を比較するためには、編集距離 [2] と言われる指標を用いて行われるが、本研究においては、編集距離の定義を拡張することにより、時系列データの動向を捉え、それを基にしてデータ間の相関関係を得ることを可能にする。これにより時系列データ間の関連性を捉え、その結果をわかりやすく言葉で説明する。

2 SAX 法を利用した時系列データの関連性の発見

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 $\bar{C} = \bar{c}_1, \dots, \bar{c}_w$ 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。

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

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。標準化された時系列にガウス分布があるとし、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)。

A Study of Linguistic Summarization based on Relevance Discovery of Multiple Time-series Data
Mizuki KOBAYASHI(g0820516@is.ocha.ac.jp),
Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
Dept. of Information Sciences, Faculty of Science, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

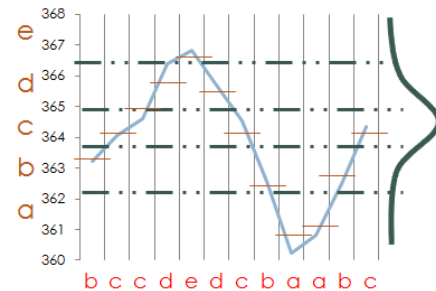


図 1: SAX 法による文字列変換

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)。

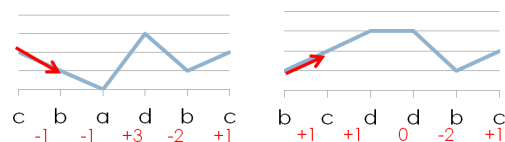


図 2: 編集距離の拡張

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の比較をし、2 つの時系列データ間の以下の 4 つの関係を取得することができる。

1. 同じ動き
-動きを示す値が全く同じ箇所 (図 3 参照)。
2. 類似の動き
-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 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 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「一」を当てはめ、その記号が全く逆の箇所(図6参照)。

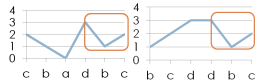


図 3: 同じ動き

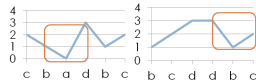


図 4: 類似の動き

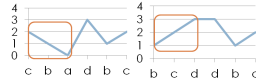


図 5: 対称の動き

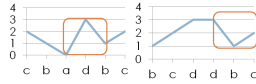


図 6: 対称に類似した動き

2.3 階層的マッチング

大局的な視点も局所的な視点も含めた分析をする場合、全体を対象に作業をすると計算時間もかかり無駄も多い。そこで、本研究では2.1節や2.2節の作業を階層的に行うことで調べる対象を限定し、マッチングにかかる計算時間を削減する。

まず、データ全体にSAXを粗くかけ、類似の動きをする箇所を探す。そこで類似していると判断された箇所の中でまたSAXをかけ、類似の動きをする箇所、対称の動きをする箇所をそれぞれ取得する。さらに類似した箇所の中だけで、また類似の動きをする箇所、対称の動きをする箇所を取得していく。この作業をあらかじめ指定した回数繰り返す。この作業が終了した後、次に対称の動きをする箇所を探す。

3 抽出された関連性を言語化

2.3節の階層的マッチングによって得られた関係性と2.2節で定義した編集距離を用い、以下のようなテンプレートに当てはめ言語化する。以下にテンプレートの一例を示す。

「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」

上記のテンプレートは、どちらも下落の動きで「似たような動きをしている」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ときに適用される。

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. 比較対象の抽出

大量にある時系列データの中でも特に関連のありそうな2つのデータを抽出するため、時系列データ間の相関係数を求め、あらかじめ設定した閾値以上の値を示すペアを取り出す。

2組の数値からなるデータ列 $(x, y) = (x_i, y_i) (i = 1, 2, \dots, n)$ が与えられたとき、相関係数は以下の(2)式で表される。

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

step 3. 比較

取り出した2つの時系列データ間で、SAXとマッチングの作業を階層的に繰り返す。

step 4. 表示

発見されたデータ間の関連性を、文章とグラフを用いて表示(図7参照)。

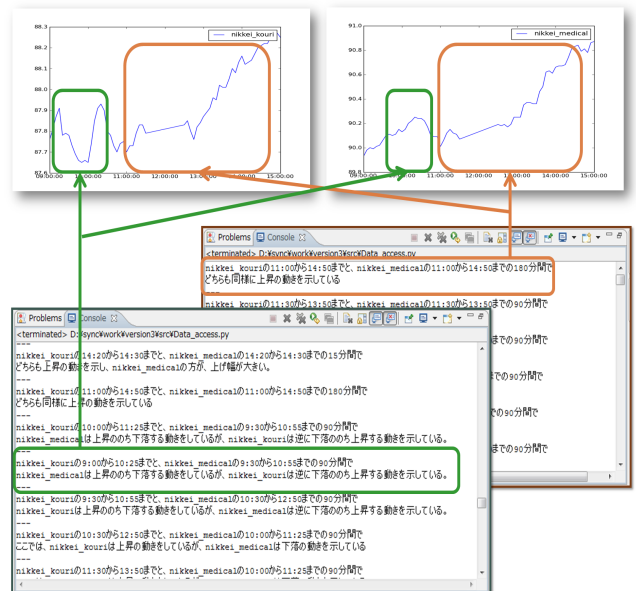


図 7: 関連性を示す言語とグラフの対応

5 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、無意味な関連性の表示を減らすため分析方法を見直すとともに、言語表現の種類をさらに増やしていきたいと考える。

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.