SAX法による局所パターン抽出を導入した時系列データの三次元可視化

理学専攻・情報科学コース
井元麻衣子

1 概要

本論文では，大規模な時系列データを大局的にも局所的にも可視化するための，三次元可視化手法 [1] を提案する．本手法ではx軸に時間，y軸に数値を割り当て，算出した類似度に従って大量の折れ線群をz軸方向に等間隔に並べて配置する．続いて，本手法では，ユーザがxz平面に垂直な視点から折れ線グラフ群全体を大局的に観察し，詳しく観察したい類似する有限本数の折れ線グラフ群を選択する．同時に，本手法はxy平面に垂直な視点から選択された折れ線グラフ群を局所的に表示する．また，SAX法 [2](Symbolic Aggregate approXimation) によって局所パターンを抽出し，それらをユーザの操作に合わせて表示する．本手法によって，大規模時系列データの全体像を眺めながら，興味のある少数の数値群を選択的に注視し，折れ線グラフ同士の相関性を発見する．

2 提案内容

2.1 概要

提案手法は，以下の4つの処理手順で構成される．

1. デンドログラムを用いた折れ線の順列化
2. SAX法を用いた局所パターンの抽出
3. 真上視点部分による折れ線グラフ全体の表示
4. 詳しく観察する特定の折れ線群の抽出，および正面視点部分による観察

本手法は図1(Left) に示すように，x軸を時間軸，y軸を数値として，折れ線グラフをz軸上に等間隔に並べて可視化する．これにより，折れ線同士の複雑な絡み合いが解消され，折れ線1本の各時刻における数値を読み取れる．また，三次元空間内に配置することで，複数の視点からグラフを観察できる．

2.2 デンドログラムを用いた折れ線の順列化

時刻 i における j 番目の数値を a_{ij} とし，j 番目の数値群によって構成される1本の折れ線を $A_j = a_{1j}, \cdots, a_{nj}$ とする．このとき本手法は，以下のいずれかの判断基準により任意の2本の折れ線間の距離を算出し，デンドログラムを生成する．そして，生成されたデンドログラムによって折れ線群を順列化する．ここでは，同一時刻に近い値を有する傾向が大局的に見られる折れ線が近くに配置されるように折れ線を並べる場合について説明する．折れ線が N 本あるとする．このとき，任意の2本の折れ線 A_j，A_k 間の距離の算出には最短距離法を用いる．ます，2本の折れ線を n 次元ベクトルとみなし，そのマンハッタン距離を算出する．

次に，$S_n(j,k)$ の値が最も小さい2本の折れ線を併合し，1つのクラスタ K_1 を生成する．ここで K_1 は，含まれている2本の折れ線の各時刻における数値の重心をその時刻での数値とし，それらにより生成される1本の折れ線とみなす．新しく折れ線を生成したことにより，折れ線の本数は N-1本となる．この操作を繰り返し，折れ線の本数を減らしながらクラスタリングを行うことにより，デンドログラムを生成し，数列群を順列化する．このとき，デンドログラムは，$S_n(j,k) \leqq S_n(l,m) \leqq \cdots$ となるように，昇順に並べる．

2.3 SAX法を用いた局所パターンの抽出

提案手法が採用するSAX法（Symbolic Aggregation approXimation）[2] は，時系列データのパターン抽出および検索のために，時系列データを文字列によって表現する手法である．具体的には図1(Right) に示すように，時間軸を等間隔に分割し，分割した時刻における折れ線の数値をアルファベットに変換し，折れ線を1つの文字列として表現する．このとき，アルファベットの出現回数が同程度になるように，y座標値の境界値を設定する．このように，折れ線を1つの文字列で表現することにより，自然言語処理分野における文字列処理，言語解析などのアルゴリズムを適用することができる．

2.4 折れ線グラフの全体表示と少数の折れ線群抽出

本手法では，xz平面に垂直な視線 Viewpoint1 と xy平面に垂直な視線 Viewpoint2 を用意し，両者を併用するこ

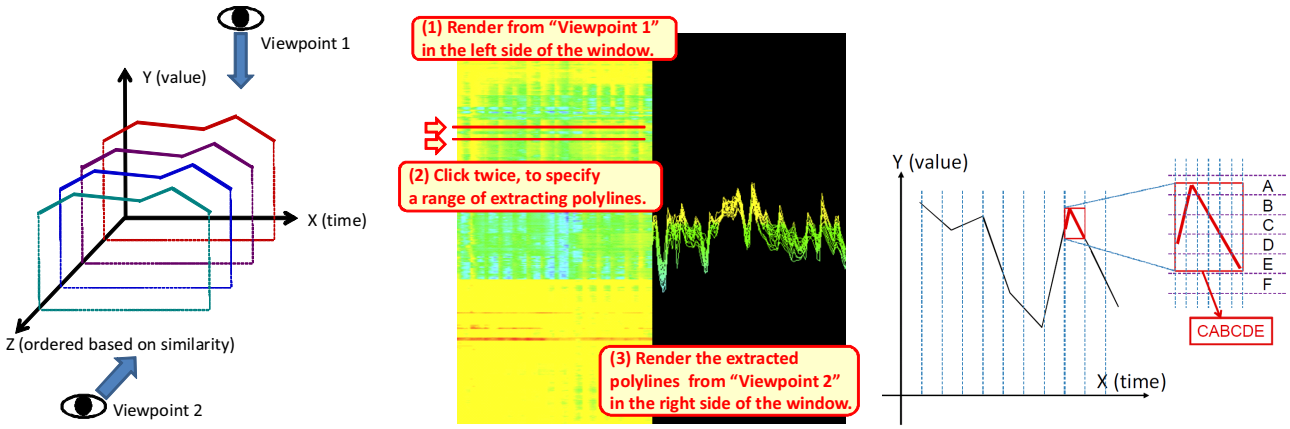

図 1: (Left) 折れ線を並べる．(Center) 少数の折れ線群の選択と表示．(Right) SAX 法を適用して折れ線を”CABCDE”と表現する．

とで大局的な可視化と局所的な可視化の相互利用を容易にする．

まず，図 1(Center) に示すように，画面左側で xz 平面に垂直な視線 Viewpoint1 により，折れ線グラフ群の全体を上から俯瞰する．このとき，折れ線の各時刻の数値を色で表現する．具体的には，y 座標値が大きければ暖色，小さければ寒色を与え，各数値の変化を色で表現するこれにより，具体的な数値として観測することのできない y 座標値や微分係数を把握する．

そして，特に着目したい折れ線群をその色分布から発見したら，その近くに視点の z 座標値を移動させ，その部分をズームアップし，クリック操作により少数の折れ線群の範囲を指定する．このとき，指定する折れ線群の本数は制限しない．この操作に伴って，画面右上では xy 平面に垂直な視線 Viewpoint2 で，クリックされた範囲に含まれる折れ線群を可視化する．

以上によって，大量の折れ線の中から注目する類似折れ線群を抽出し，その関連性や差異を観察できる．折れ線グラフ全体を上から俯瞰することで，重要な意味をもつと考えられる折れ線を見落とすという問題点は解決されると考えられる．

3　適用事例

図 2 は日本の気象観測システム AMeDAS(Automated Meteorological Data Acquisition System) が 2006 年 1 月に観測した全国 916 地点の気温データを SAX 法を適用して可視化した例である．本研究では時系列データを 4 時間ごとに区切って，各々の区間に 7 種類の文字を割り当てた．そして，1 日の気温変動に相当する 6 文字を 1 つの文字列とし，出現回数が 300 以下であった文字列を異端パターンとして抽出し，その局所パターンの 1 つを紫の枠で表示した．ここで興味深い点として，真上視点からクリックした 2 つの範囲 (以下,「グループ赤」「グループ青」と呼ぶ) は，全体的な気温の時間変動が非常に類似しており，また，左側の黄色の四角形内部には，抽出されたパターンが多く存在する．一方で，右側の黄色の四角形内部には，グループ青には抽出パターンがないが，グループ赤には抽出パターンが多く存在する．正面視点で描画してみると，グループ赤は 1 日の気温変動が他とは異なっており，日中に気温が上がらなかったことがわかる．

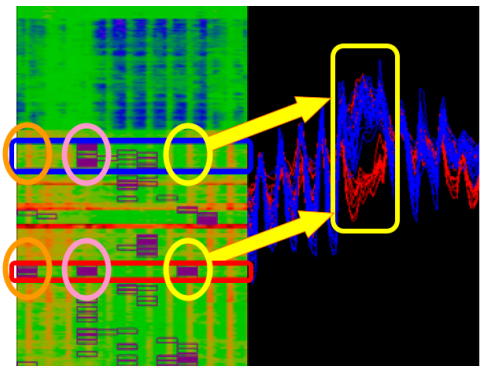

図 2: SAX 法を適用して可視化した例．

このように本手法を用いることで，まず大局的に時系列データを眺めて興味深い類似データ群を抽出し，続いてそれらを局所的に比較することが容易になる．

4　まとめ

本論文では，時系列データを表現する折れ線群を三次元空間 (xyz 空間) に配置し可視化する一手法を提案した．

参考文献

[1] M. Imoto, T. Itoh, A 3D Visualization Technique for Large Scale Time-Varying Data, *14th International Conference on Information Visualisation (IV10)*, pp. 17-22, 2010.

[2] J. Lin, E. Keogh, S. Lonardi, B. Chiu, A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, *8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2003.