M-053

都市環境データのクラスタリングに関する一考察

A Study of Clustering for Urban Environmental Data

大野 貴弘† テープウィロージャナポン ニワット‡* 戸辺 義人‡*
Takahiro Ono Niwat Thepvilojanapong Yoshito Tobe

1. まえがき

近年，複雑化した都市環境を把握するため，細粒度センサネットワークが展開されてきている．そして，細粒度のデータを収集することで，道幅や街路樹と気温の相関などの都市環境情報が抽出できるようになった．しかし，それらの情報は観測者がデータを目視することによって抽出するため，センサネットワークが拡大し，データ量が増大した場合，解析コストが膨大となり，情報抽出が困難となる．そこで本研究では，細粒度センサネットワークから得られる大量のデータを効果的にクラスタリングし，データ解析をサポートする方法について考察する．本稿では，2007 年夏に実施された実験(UScan[2])のデータを用いて，気温データから得られる情報に焦点を当て，議論を進める．

2. 関連研究

大量データを効率的に解析する手法としてクラスタリング[1]がある．クラスタリングは，データを一定のルールにしたがって分類し，解析を行うので，処理コストが少なく，大量データの解析に適している．クラスタリングの適用分野としては，パターン解析やデータマイニング，文書検索などがある．ネットワークやデバイスが発達し，カメラや気象センサなどから収集できるデータ量が増加した現代において，クラスタリングは，重要なデータ解析技術の 1 つである．

クラスタリング手法は大きく 2 種類(階層クラスタリングと空間クラスタリング)に分けられる．階層クラスタリングには，クラスタ間の距離を測る代表的な 2 つのメソッドがあり，Single-link と Complete-link と呼ばれる．Single-link では，クラスタ中の要素同士の最小距離を使用し，クラスタ間の距離を測るのに対し，Complete-link では，最長距離を使用する．Single-link では，クラスタが大きい場合，要素間とクラスタ間の最小距離の差が大きくなり，うまく計算できないので，一般には Complete-link の方が使われることが多い．空間クラスタリングでは，多次元空間上にデータを配置して処理を行う．クラスタ間の距離は，ユークリッド距離を用いることが多い．実際，グラフ理論を用いたクラスタリングでは，データの要素同士の最小距離を基にエッジを作り，エッジの最大距離でクラスタリングを行う．

本稿では，空間クラスタリングを用いてデータ解析手法を検討する．これは，空間クラスタリングの方が，直観的にデータの分布を確認することができ，データの分布状況も視覚的に認知できるため，データから情報を読み取るのに，適しているからである．

† 東京電機大学大学院 工学研究科
‡ 東京電機大学 未来科学部
* 科学技術振興機構 CREST

3. クラスタリング手法

3.1 歪みと尖りの定義

・歪み

気温のデータが基準時間を中心として，右側に歪んでいるのか，左側に歪んでいるのかを表す指標である．ここでの基準時間とは，ある時系列データの開始時間と終了時間の中間となるような時刻である．歪みの値は，時系列データを式(1)で変換して求める．右に歪んでいる場合(基準時間よりも後半の方が温度が高い場合)，歪みの値はプラスに，左に歪んでいる場合は，歪みの値はマイナスになる．

$$\text{歪み} = \frac{1}{n}\sum_{i=1}^{n} T\_t_i * \alpha_i \quad \cdots \quad (1)$$

・n：データ数
・T_t：時刻 t における気温データ
・α：重み [-n/2 < α < n/2]

・尖り

気温のデータの変化量を求め，その絶対値の平均値をグラフの尖りと定義．気温の上昇と下降が激しいグラフほど尖りの値は高くなる．下記にその式(2)を示す．

$$\text{尖り} = \frac{1}{n}\sum_{i=1}^{n-1} \left| T\_t_{(i+1)} - T\_t_i \right| \quad \cdots \quad (2)$$

3.2 クラスタリング領域

3.1 で定めた歪みと尖りで，気温データを処理し，横軸を歪み，縦軸を尖りとして 2 次元座標に配置する．また，クラスタリングのために 2 次元空間を下記のような定義で 4 分割する．本稿では，データが下記の定義領域にどのように分布しているかを見ることにより，データ解析を行う．

・領域の定義

領域 I：歪み ＞ 0 かつ 尖り ≧ 1
領域 II：歪み ≦ 0 かつ 尖り ＞ 1
領域 III：歪み ＜ 0 かつ 尖り ≦ 1
領域 IV：歪み ≧ 0 かつ 尖り ＜ 1

4. データ解析

4.1 対象データ

(1) 期間：2007 8/21 – 8/27 (10:00-20:00)
(2) 場所：東京電機大学神田キャンパス周辺
(3) データ：気温

本稿では，2007 年度夏に実施した実験(UScan)のデータを用いる．UScan とは，細粒度センサネットワークを活用することで，詳細な都市環境を把握し，ナビゲーションシステムなどに活用させ，安心・安全な社会を目指すプロジェクトである．図 1 のエリア（東京電機大学神田キャンパス周辺）では uPart を約 150 個配置し，気温データを取得した．都市においては，環境が複雑化しているので，センサ設置地点が近くても，図 2 のような様々なタイプの気温グラフが得られる．

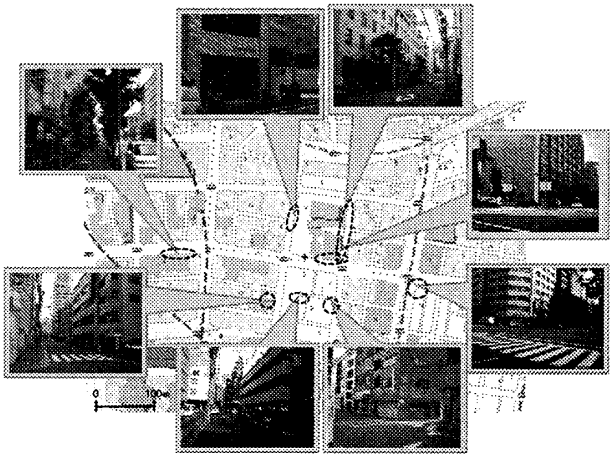

図 1. 実験エリア概要

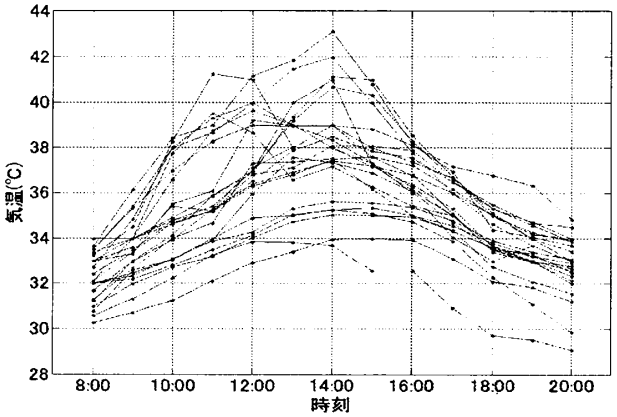

図 2. 2007 年 8 月 22 日 気温グラフ

4.2. クラスタリング結果

対象データを 3.1 の式で計算し，二次元座標にマッピングすると，図 3 のような結果が得られる．

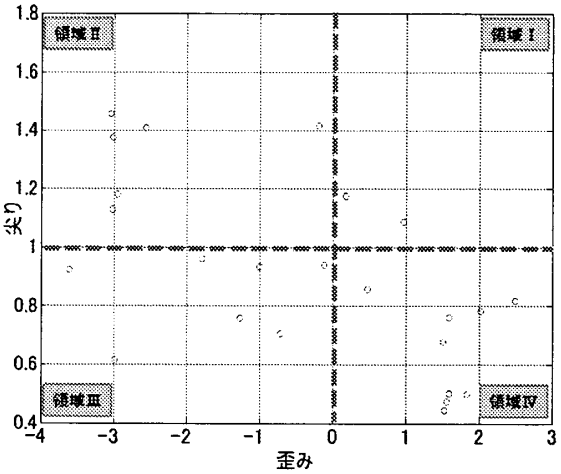

図 3. 歪みと尖りの分布の例 (2007 年 8 月 22 日)

そして計算結果に対し，3.2 のクラスタリングを実行しデータ分布をみると，図 4 のようなグラフが得られる．

図 4 は，対象データに対し，領域（Ⅰ・Ⅱ・Ⅲ・Ⅳ）に占めるデータの割合を示したものである．日によってデータの偏り具合が変化していることがわかる．

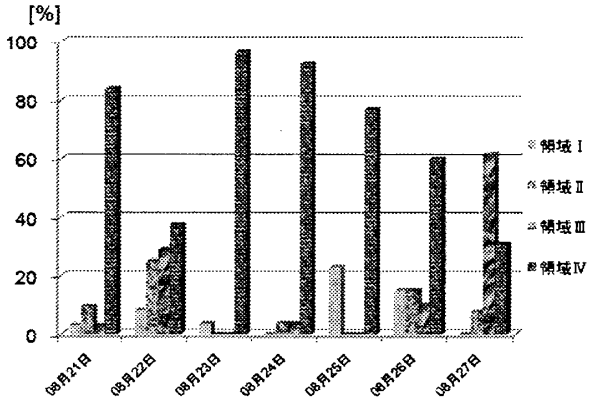

図 4. 歪みと尖りによるデータの偏り

4.3. 結果の考察

終日晴天であった 8 月 22 日のデータを基準に相対的に分析する．まず，領域Ⅳの割合が一番高い 8 月 23 日と比較する．分布状況から，22 日のデータより，23 日のデータの方が右にシフトしていることがわかる．右にシフトしているということは，気温が 14:00 以降の方が高くなっているということである．実際，23 日の午前中は雷を伴った雨が降り，午後から晴れてくるという午後から気温が上昇する天気であった．他の，21 日，24 日の午前中も，雨を伴うほどではないが，雲の影響で日照量が減少し，午前の気温が下がったため，23 日と同じ傾向を示している．27 日は 23 日のグラフと逆の傾向を示している．実際，27 日の天気は午後から曇りである．このように，気温データをクラスタリングすることで，天気の傾向を知ることができた．

今回は，データを 4 つの領域に分けるクラスタリングを実行し，データ分布を分析しただけであるが，雨と曇りの識別，晴れとうす曇りの識別も，データ分布解析により可能となる．実際に，図 4 を見ると，21 日と 23 日の傾向は同じに見えるが，歪みと尖りデータ分布を比較すると 23 日のデータの方が，より右にシフトしていることが確認できた．他にも，設置地点と気温の相関など都市環境データは，さまざまな情報を含んでいるので，抽出したい情報に合った重みを設定すれば，さらに効果的なデータ解析が実行できると考えられる．

5. まとめ

本稿では，2007 年夏の実証実験のデータ解析を基に，クラスタリングに関して考察を行った．今後は，今回得られた知見をもとに，新しいクラスタリング手法の検討と実装・評価を進めていく．

参考文献

[1] Jain, A. K., Murty, M. N. and Flynn, P. J., Data clustering: a review, ACM Computing Surveys (CSUR), 31(3), 264-323, 1999.

[2] UScan http://uscan.osoite.jp/