ランキングデータにおけるクラスタリング手法の提案

金森弘晃* 松田眞一†
E-Mail: matsu@nanzan-u.ac.jp

本論文は，ランキングデータのような複数の時系列データに対して新たなクラスタリング手法を提案するものである。sliding window technique を複数時系列用に拡張した上でクラスタリングに必要な距離を考案した。結果として，視覚的に似ていると感じるデータがクラスタを構成し，ランキングデータの細かな性質を見極めることができた。

1 はじめに

時系列データの中からなんらかの特徴を導き出す研究はパターン抽出法として重要なものである。部分時系列クラスタリング (井手 [3], Das et al.[1]) はそのような方法の一つであるが，この方法は 1 本の長い時系列データを多数の部分時系列に分け，それらをそれぞれ独立なベクトルとしてクラスタリングしようというものである。このクラスタリングによってできたクラスタの平均を代表パターンとして取り出すことが目的である。

しかし，この方法はあくまでも 1 本の時系列上で連続して似たような反応が起こる場合しか扱っておらず，2 本以上の時系列があった場合や連続して反応が起きない場合のクラスタリングまでは扱っていない。

本論文ではこれまで扱われなかった単反応複数時系列のクラスタリング手法を提案する。特にポイント変化のあるランキングデータにおける変動パターンをつかむ手法を考える。したがって，以下では時系列データの単位を便宜上ポイントと呼ぶことにする。

2 アプローチ

基本方針として 2 つの時系列データ間に適当な距離を定めた上で一般のクラスタリング手法を用いることとする。その距離の算出方法であるが，時系列の長さにばらつきがある場合も考えたい。例えば，あるデータは 18 回しかランクインしなかったため 18 回分のデータしかないが，あるデータは 26 回分のデータがあるといった場合，時系列の長さに差があるものを比較する新たな方法が必要となるからである。

3 ポイント補正方法

2 つの時系列データ間距離を計算する際に，前処理としてそれぞれのデータについてポイントの補正が必要となる場合がある。もしそれぞれの時系列データのポイントに全体的な大小の差があれば，そのままクラスタリングを行うと「高ポイントのクラスタ」「中ポイントのクラスタ」「低ポイントのクラスタ」の 3 つのクラスタに分かれてしまうのは明らかで，変動パターンをつかむという本論文の目的とは一致しない。補正には平均ポイントや

*南山大学数理情報学部数理科学科
†南山大学数理情報学部情報システム数理学科

最高ポイントなどそれぞれの時系列データに関する何らかの特性値を計算し，その特性値で各回のポイントを割ることによって全ての時系列データのポイントをそろえる方法がある。ポイント補正方法として以下の 3 種類を考える。

1. 最大値で割る
2. 平均値で割る
3. 上位値の平均で割る

3.1 最大値で割った場合

最大値で割った場合のクラスタ平均ポイントの例を**図 1・図 2** に示す。（以下どの場合も前者が元データでプロットしたもので後者が実際に割った後の値でプロットしたものである。また，図中の黒線は全体平均を表し，それ以外はクラスタ平均を表す。）

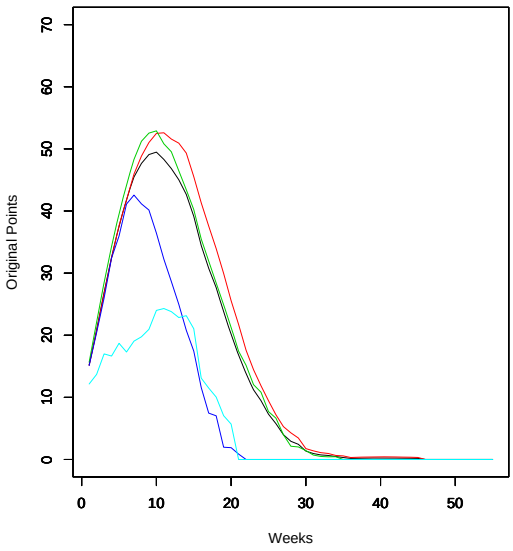

図 1: 最大値で割った場合の平均ポイント（元データでプロット）

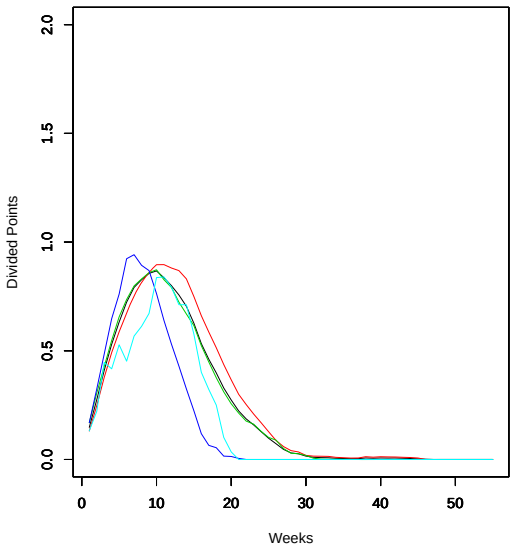

図 2: 最大値で割った場合の平均ポイント（実際に割った後の値でプロット）

利点は最大値で割ることによってポイントの大小をある程度無視したクラスタリングができることである。**図 1** を見ると，他の方法に比べどのクラスタも最高ポイントがまとまっていることが分かる。一番高いものでも 50 ポイントを少し超えたぐらいで，明確な「高ポイントのクラスタ」が存在していないことが分かる。

一方欠点は，ピークタイミングによる群分けはされにくいことである。最大値で割った後の値の**図 2** ではピークタイミングに差があり，早いものは 7 週前後，遅いものでは 12 週前後にピークを迎えているが，これはピークタイミングによって群分けされたというよりは時系列の長さによって群分けされていると見る方が自然である。なぜならデータ数の少な

いクラスタを除きどれも逆U型・逆V型といった「ピーク維持期間」や，ランクイン後すぐに最大値を記録するのか，ランクイン後末期に最大値を記録するのかという傾向があまり表れていない。また，**図 1** において上昇スピードがどのクラスタもほぼ同じであり，違いが分かりづらい。他には，最大値という 1 つの値を使い計算するのでこの値が少し違うだけで全体が大きく左右されやすいという面を持っている。

3.2　平均値で割った場合

平均値で割った場合のクラスタ平均ポイントの例を**図 3**・**図 4** に示す。

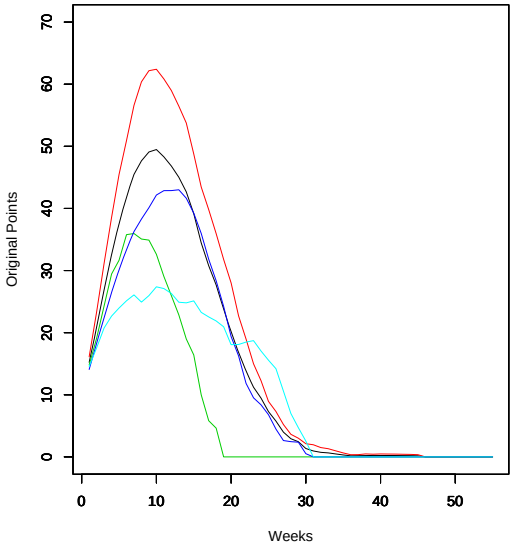

図 3: 平均値で割った場合の平均ポイント（元データでプロット）

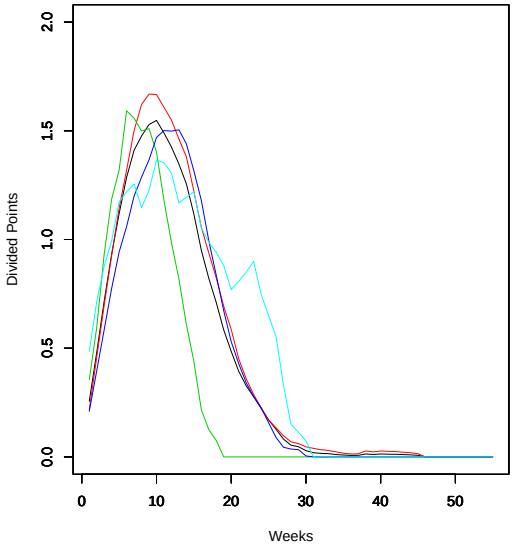

図 4: 平均値で割った場合の平均ポイント（実際に割った後のプロット）

利点はピークタイミングを考慮したクラスタリングができることである。**図 4** を見ると時系列の長さがほぼ同じであるが，ピークタイミングや上昇スピードが異なるクラスタが存在する。このようにそれぞれのデータの平均ポイントが異なるため，割った際の最高ポイントもランク変動パターンによって違いがあり，全てを加味したクラスタリングができる。

一方，欠点はポイントの全体的な大小によって群分けされやすいことである。データの補正後最大値（補正前最大値/平均値）の最高は今回解析したデータでは 2.24 であるが，補正後最大値の最低は 1.13 と 2 倍近い差がある。さらに，補正前の各データの最大値と補正後の各データの最大値の間には 0.72 という正の相関があることが分かった。これでは 2 つの時系列データ間のポイント差を補正する方法としては不十分であり，結果的にポイントの大小で分かれてしまうことになる。また，低ポイントが長く続くとその結果に左右されてしまう。ピークポイントが普通であっても低いポイントのままランクインし続ける場合，平均値がとても小さくなり，平均値で割った補正後の最高ポイントが他のデータよりはるかに高くなってしまう場合も出てくる。

3.3　上位値の平均で割った場合

上位値の平均で割った場合のクラスタ平均ポイントの例（ここでは上位 5 値を用いた）を**図 5・図 6** に示す。

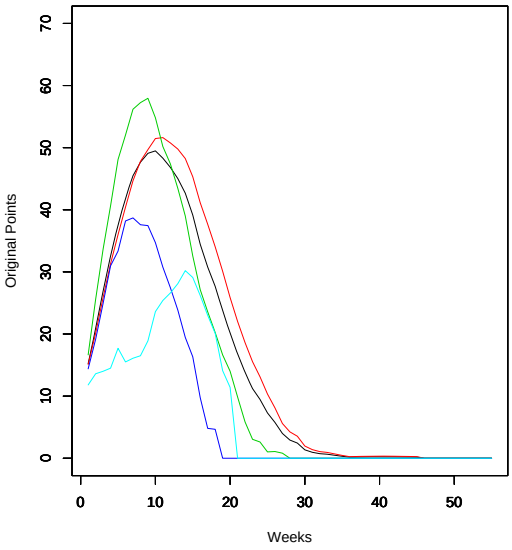

図 5: 上位値の平均で割った場合の平均ポイント（元データでプロット）

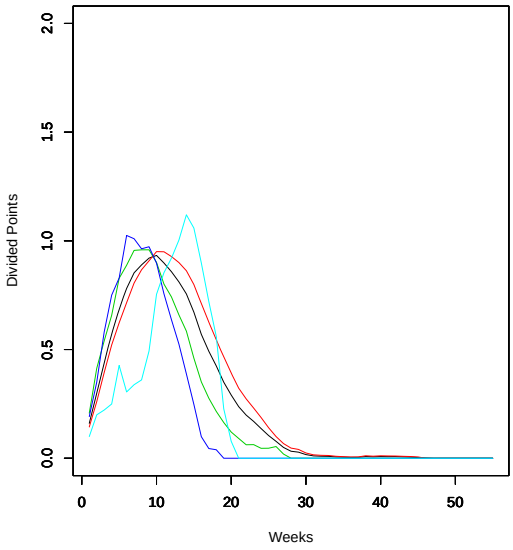

図 6: 上位値の平均で割った場合の平均ポイント（実際に割った後のプロット）

この上位値の平均で割る方法に明確な利点はない。しかし**図 5** を見ると**図 1** ほどではないがある程度ポイントの大小を無視したクラスタリングができていることがわかる。実際にこの方法による補正前の各データの最大値と補正後の各データの最大値の相関係数は 0.14 であり，ほとんど影響はないことが分かる。また，**図 6** においてはピークタイミングが同じであるがランクイン期間の異なるクラスタが存在するなど最低限の群分けができている。1 つの最大値やポイントの全体的な大小によって補正後の値が大きく変わってしまうようなこともなく，安定したクラスタリング結果が得られる。

3.4　まとめ

最大値や平均値で割る方法はどちらも利点はあるが欠点が非常に大きい極端な方法であった。その上，補正後の値が一つの最大値によって左右されたり，低ポイントが続くことによって正確な補正ができない例もあった。一方，上位値の平均で割る方法は両者の間をとったようなものではあるが，この方法が最も欠点の少ないものである。今回は上位 5 値の平均を使ったが，上位のいくつ (上位 x 値) まで使うかによって結果も変わってくる。つまり何を求めたいか，どんなクラスタリングをしたいかによってこの値を変えればこの方法の実用性はさらに高まる。x を 1 に近づければ最大値で割る方法に近くなり，逆に x の値を

増やせば平均で割った場合に近くなる。なお，今回のように単反応で変動が激しくないものであれば上位 5 値ぐらいが妥当であろう。

4　2 つの時系列データ間の距離行列

2 つの時系列データ間の距離の算出にはまず，井出 [3] の sliding window technique を使っていくつかのパートに分ける必要がある。井出 [3] では対象となる時系列が 1 つだったため一方のみをスライドさせる方法であったが，ここでは 2 つの時系列の比較であるため両方のデータをスライドさせる点で違いがある。

次に 2 つの時系列データ $\boldsymbol{a}$, $\boldsymbol{b}$ のそれぞれのパート間の距離を求める。sliding window technique の固定窓幅 ω を 15 とすると $\boldsymbol{a}$ のパートの 1 つ目 $A_{1r}(r = 1, 2, ..., \omega)$ は a_1 から a_{15} まで，2 つ目 $A_{2r}(r = 1, 2, ..., \omega)$ は a_2 から a_{16} まで，3 つ目 $A_{3r}(r = 1, 2, ..., \omega)$ は a_3 から a_{17} までという具合に部分時系列を作っていく。最終的には長さ n の 1 つの時系列データから $n - \omega + 1$ 個の部分時系列が生成されることになる。$\boldsymbol{a}$ の部分時系列 $\boldsymbol{A}$ と $\boldsymbol{b}$ の部分時系列 $\boldsymbol{B}$ を

$$\begin{cases} \boldsymbol{A}_i = (A_{i1}, \ldots, A_{i\omega})' \quad (i = 1, 2, ..., n - \omega + 1) \\ \boldsymbol{B}_j = (B_{j1}, \ldots, B_{j\omega})' \quad (j = 1, 2, ..., m - \omega + 1) \end{cases} \tag{1}$$

とおくことにする。ただし，n, m はそれぞれ $\boldsymbol{a}$, $\boldsymbol{b}$ の時系列の長さである。

最後にこの 2 つの時系列データ間の距離行列を計算するのであるが，それは部分時系列同士のユークリッド距離によって求めることとする。すなわち，$d(\boldsymbol{x}, \boldsymbol{y})$ を同じ長さの 2 つのベクトル $\boldsymbol{x}$, $\boldsymbol{y}$ 間のユークリッド距離として，2 つの時系列データ間の距離行列 $X = \{X_{ij}\}$ を以下のように定義する。

$$X_{ij} = d(\boldsymbol{A}_i, \boldsymbol{B}_j) \tag{2}$$

5　距離行列の意味

距離行列 X は時系列データ間の全ての一部分どうしの距離をあらわしたものである。このような操作が必要である理由はデータによってピークのタイミングが違うからである。例として**図 7** を挙げる。縦軸はポイント，横軸は週単位の時間である。この 2 つの時系列データは最初の上昇スピードは違うがその後の推移はかなり似ている。これはただ単に急上昇した時期が少しずれていただけと考えるのが自然であり，データ間の距離は近くなるのが理想と考えられる。そこでこの距離行列は全ての一部分どうしの距離を計算しているため，2 つの時系列データ間で最も距離の近くなった週どうしを選んで総合的な距離として適用することができる。

6　2 つの時系列データ間距離

2 つの時系列データ間距離 D は以下の式で定義する。

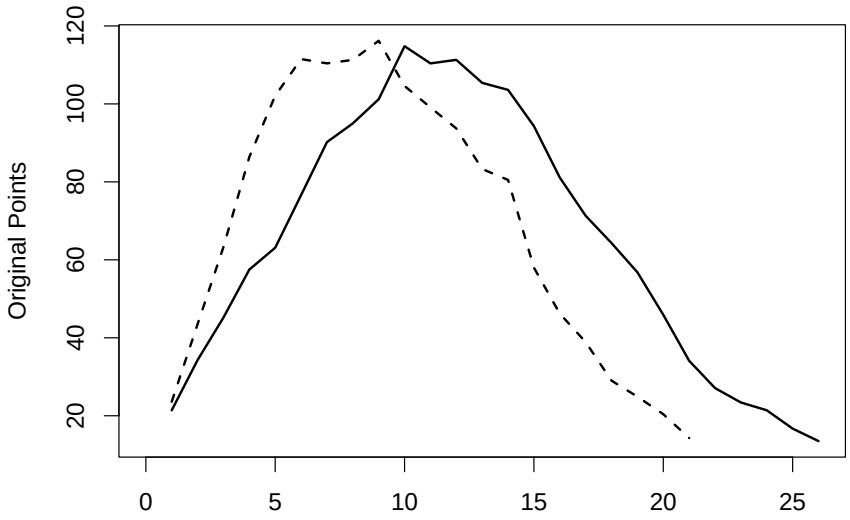

図 7: タイミングの差

$$D = \frac{1}{2}\left(\frac{\sum_{k=1}^{n-\omega+1} \min X_{kj}}{n-\omega+1} + \frac{\sum_{k=1}^{m-\omega+1} \min X_{ik}}{m-\omega+1}\right) \tag{3}$$

すなわち，まずそれぞれの行方向について最小値をとり，その最小値の平均をとる。同様にそれぞれの列方向についても最小値をとり，その平均をとる。最後に行方向の平均値と列方向の平均値の平均を 2 つの時系列データ間距離とする。

前節でずれを考慮して一番近くなる部分時系列を選ぶと述べたが，単純に距離行列の全体の最小値ではうまくいかなかった。全体の最小値は細かな凹凸の偶然の一致を表すことが多いのが原因でないかと思われる。そのため上記のようにずれを考慮して最小値を探すものの全体としての曲線の当てはまりも考慮してその平均をとるという方法を考案した。（距離行列の全平均では全然当てはまっていない部分時系列どうしも考慮されることになり，うまくはいかない。）行方向の平均と列方向の平均でさらに平均を取る理由は対称性を保つためである。

7　クラスタリング

前節の距離を元にクラスタリングを行う。その際最長距離法を用いてクラスター分析をすることにする。（クラスター生成方法は様々あるが，実例で問題が生じなかったので本論文ではその点は追求しなかった。可能性としてはウォード法の方が向いている場合もあると思われる。）

8　固定窓幅 ω

固定窓幅 ω は一般的にウィンドウサイズとも呼ばれているもので，長い時系列データをどのくらいの幅で分けていくのかを決めるものである。井手ら [2] では固定窓幅 ω の設定について,「変化点の発生間隔が ω より小さいと探知は難しいことに注意する必要がある。」と述べている。本論文では単反応を考えているので，変化点の発生間隔とは時系列の長さそのものであり，固定窓幅 ω より短い時系列データを解析に用いないため今回の手法では問題がない。結論として固定窓幅 ω が長すぎると解析可能なデータ数が減ってしまうことから，できる限り短い幅が望ましい。しかし，固定窓幅が短ければ良いというわけでもない。**表 1** はある時系列データ $\boldsymbol{a}$（時系列の長さ 40）と，時系列データ $\boldsymbol{b}$（時系列の長さ 32）の距離行列の一部を固定窓幅 ω 別に並べてみたものである。

$\omega = 10$	$\omega = 11$	$\omega = 12$	$\omega = 13$	$\omega = 14$	$\omega = 15$
1.28	1.46	1.64	1.84	2.08	2.36
1.41	1.57	1.75	1.97	2.24	2.42
1.53	1.69	1.90	2.14	2.32	2.47
1.66	1.86	2.09	2.25	2.38	2.51
1.85	2.07	2.21	2.33	2.45	2.52
2.07	2.20	2.31	2.42	2.48	2.55
2.20	2.31	2.41	2.46	2.53	2.54
2.31	2.41	2.46	2.51	2.51	2.51
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
1.78	1.80	1.82	1.83	1.84	1.84
1.70	1.72	1.73	1.74	1.74	1.74
1.60	1.62	1.63	1.63	1.63	1.63
1.49	1.50	1.51	1.52	1.52	1.52
1.34	1.38	1.39	1.40	1.40	1.40
1.20	1.24	1.26	1.26	1.26	1.27
1.06	1.09	1.11	1.12	1.12	1.13
0.88	0.92	0.94	0.95	0.95	0.96

表 1: 固定窓幅 ω の違いによる距離行列の変化

表の上部は $\boldsymbol{a}$ の序盤と $\boldsymbol{b}$ の終盤との距離で，表の下部は $\boldsymbol{a}$ の終盤と $\boldsymbol{b}$ の終盤との距離である。本来ならば時系列データ同士の序盤は序盤と，終盤は終盤との距離が近くならなければならないのであるが，$\omega = 10$ の場合序盤と終盤の距離が 1.28 という近い距離になってしまっていることが分かる。この傾向は時系列データが長いほど顕著で，行方向と列方向の最小値を 2 つの時系列データ間距離算出のもとにしている今回の手法ではできる限り避けたい現象である。そこでこの例以外でも固定窓幅 ω の値を変えて実験してみたところ $\omega = 15$ あたりでどの時系列データ間も安定してきたため，今回の解析では固定窓幅を 15

で行うことにする。

この固定窓幅はどのような時系列データであるかに大きく依存する可能性が高く，解析ごとに最適と思われる値を探す必要が生じる。ただ今回の検討により単反応であれば時系列データのうち最も短いものの長さに近い値で概ね実用的なものとなると思われる。いくつかの例外的に短い時系列データのために ω を小さくしすぎないことが重要である。

9　音楽ランキング解析例

例として Top Hits Online[4] の 2003 年の音楽ランキングを解析する。今回の解析では固定窓幅（週数幅）ω を 15 週，ポイント補正方法は上位 5 値の平均で割る方法をとった。解析で得られた**図 8** のデンドログラムを左から大きく 5 つのクラスタに分け考察する。

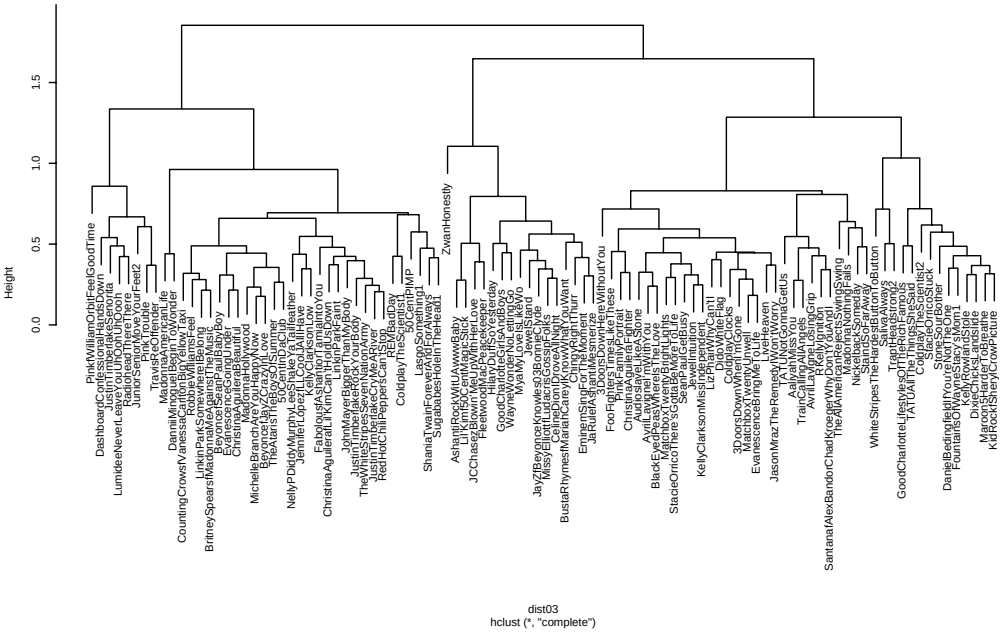

図 8: 2003 年デンドログラム

図 9 が 5 つのクラスタの平均を示したものである。（黒線は全体平均である。）

- 第 1 クラスタ （緑線）
  全ての曲が 21 週以下とランクイン期間が短い曲がこの群を占める。ヒットタイミングも早い上にピーク後の下降スピードも急激である。

- 第 2 クラスタ （青線）
  第 1 クラスタに似ているがこのクラスタの登場週数を見てみると 18 週から 32 週まで幅広く，特にランクイン期間が短いというわけではない。しかし第 1 クラスタに近いスピードで上昇し，ヒットタイミングが早いのが特徴である。

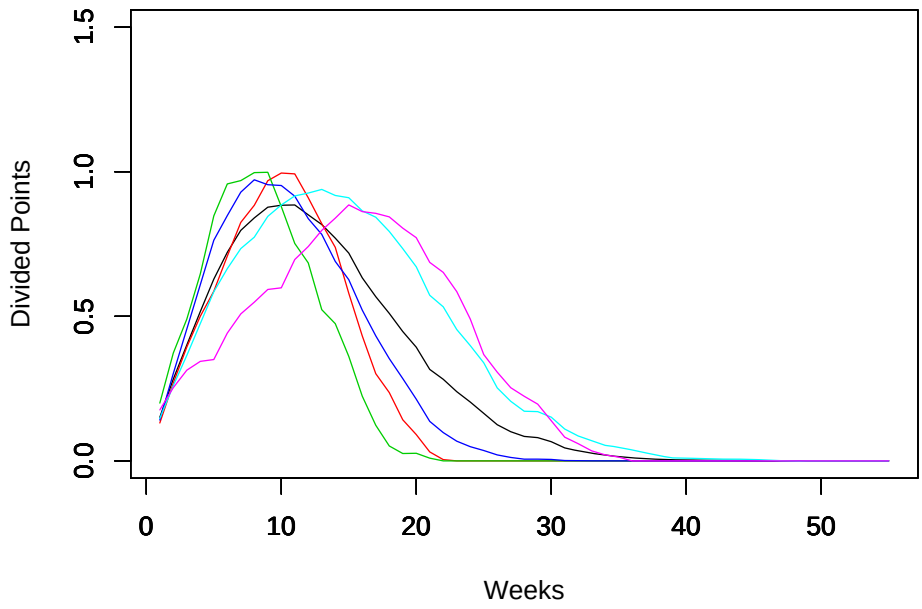

図 9: 2003 年クラスタ平均

- 第 3 クラスタ （赤線）
  このクラスタも第 1 クラスタ同様に全ての曲が 22 週以下とランクイン期間が短い。しかし，ランクイン期間が短いとどうしても早くなりがちなヒットタイミングがこの第 3 クラスタに関しては特にそういった特徴はない。むしろこのクラスタはピーク後の下降スピードが急激で目立っている。

- 第 4 クラスタ （水色線）
  このクラスタは 20 週から最高は 46 週まで幅広い曲が属すロングヒットのクラスタである。ロングヒットであるがどの曲も序盤に上昇し，昇降を繰り返すのではなくわりとしっかりしたヒットの山を持ち，緩やかに下降していく曲のクラスタである。

- 第 5 クラスタ （ピンク線）
  このクラスタの大きな特徴は序盤のポイント上昇が他のどのクラスタよりも緩やかなことである。曲のピークが 20 週前後の曲が多く，それまではわりと激しく昇降を繰り返しながら上昇していく。上昇が緩やかな反面，ピーク後は一気に下降しランク外になる曲ばかりである。

9.1 まとめ

今回の例では大きく 5 つのクラスタに分けることができた。第 1，第 2，第 3 クラスタは良く似ていて違いが分かりづらいが，単にランクイン期間が短い第 1 クラスタの「短期間

ヒット群」に対し，第 2 クラスタはランクイン期間が短いわけではなかったことから「ヒットタイミング早期群」と呼べる。また，第 3 クラスタはランクイン期間が短い曲群にしてはヒットタイミングが少し遅めで，下降が急激なため「ヒットタイミング末期群」とそれぞれ現行の手法でとても細かい特徴まで読み取れていることが分かった。他にも第 4，第 5 クラスタのようにどちらも 30 週前後のロングヒット曲が多い中でヒットタイミングによるクラスタリングができていた。第 4 クラスタはヒットがしっかりしていた「ロングヒット群」であり，第 5 クラスタは上昇スピードにおいては緩やかであるが一気に下降する「下降急激群」であった。

この他の年についても同様に解析を行ったところ，いずれの年もまず「短期間ヒット群」のような変動スピードの速いクラスタが大きく分かれた。逆にランクイン期間の長い曲についてはそれほどまとまったクラスタにはならなかった。そのため，**図 9** のようなクラスタ平均の曲線を代表と呼ぶ場合には注意を要する。すなわち，今回のように $\omega = 15$ の場合は 15 からその 2 倍の 30 くらいまでをランクイン期間でうまくクラスタリングする結果となり，序盤にピークを迎えるか，終盤にピークを迎えるかといった特性もそれぞれのクラスタに現れていたが，長めの上昇や下降に対する特性にはばらつきが大きかった。したがって，クラスタ平均からピークタイミング付近の特徴とすその方の特徴を抽出する場合，その安定性に違いがあることを理解することが重要であり，個別のデータのプロットにも注意すべきである。

10　おわりに

本論文の目的はランキングデータにおける変動パターンをつかむクラスタリング手法の提案であった。変動パターンという点に関しては解析結果にも反映され良い手法ができていた。ランクイン回数やピークタイミングが違っても似ている変動をしたものについてはまとめることができ,「視覚化した場合に似ているものをまとめる」という大きな目標は満たしていた。

参考文献

[1] G. Das, K.-I. Lin, H. Mannila, G. Renganathan, and P. Smyth. (1998): Rule discovery from time series, Proc. the Fourth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.16-22.

[2] 井手剛, 井上恵介 (2004)：非線形変換を利用した時系列データからの知識発見, 第 4 回データマイニングワークショップ, 日本ソフトウェア科学会データマイニング研究会, 研究会資料シリーズ ISSN 1341-870X, No.29, pp.1-8. (`http://spinglass.hp.infoseek.co.jp/`)

[3] 井手剛 (2006)：部分時系列クラスタリングの理論的基礎, 第 20 回人工知能学会全国大会予稿集, 2A1-2. (`http://spinglass.hp.infoseek.co.jp/`)

[4] Top Hits Online：`http://www.tophitsonline.com/`.