

動的クラスタリングにおけるクラスタの変化分析

米田 一樹^{1,a)} 松井 藤五郎² 武藤 敦子¹ 森山 甲一¹ 犬塚 信博¹

概要: データ分析の手法の一つとしてクラスタリングがある。マーケティングにおいて、顧客の購買傾向を発見するためにクラスタリング手法が用いられる。その際、クラスタの変化を見ることで購買傾向の変化が見ることができるため、クラスタの変化分析は重要とされている。Spiliopoulouらは、連続的に取得したデータに対して指定した期間内のデータを混ぜてクラスタリングを行うことを考え、データの同一性に基づいたクラスタ一致率を定義し、クラスタの変化分析を行った。しかし、従来研究では前の時刻を基準とした前向きクラスタ変化の分析しか行っていないため、クラスタ変化の正確な分析が出来ていなかった。また、同じラベルを持つ事例でも異なる事例として扱われてしまう。そこで本研究では、ラベルの同一性に基づいたクラスタ一致率を定義するとともに、今の時刻を基準にした後ろ向きクラスタ変化の分析を追加して行うことで、同じラベルを持つデータを含む動的クラスタリングにおけるクラスタ変化の正確な分析を実現した。

キーワード: クラスタリング, 時系列分析, 変化分析, 動的変化

Analysis of Cluster Change in Dynamic Clustering

1. はじめに

データマイニング手法としてクラスタリングが存在し、データに対して一定の基準からグループ分けを行いグループごとの特徴の発見が出来る。そして、データを時系列ごとにクラスタリングを行い、クラスタ変化を見ることで、クラスタとしての傾向の変化を捉えることが可能になる。そのため、クラスタ変化に関する研究は数多くあり、重要とされている。

Spiliopoulouら [1] は連続的にデータが取得される環境において、指定した期間内のデータを混ぜてクラスタリングを行い、データの同一性に基づいて定義したクラスタ一致率を用いて、クラスタの変化を定義化し、クラスタの変化を分析する研究を行った。しかし、MONICでは前の指定した期間のクラスタを基準としたクラスタ変化を研究しており、前の基準からクラスタがどう変化したかだけしか見ておらず、クラスタの変化を正確に調べるには不十分で

ある。また、データの同一性によるクラスタ一致率を定義し、クラスタの変化を分析したが、同じラベルを持つデータが繰り返し出現した場合、同じラベルを持つデータを含むクラスタが変化していくのを見ることが出来ないため、適切な変化分析をすることが出来ない。

そこで、本研究ではデータラベル数の変化が少なく、同一時刻にデータが取得される環境における動的クラスタリングでのクラスタの変化を見ることを目的に、時点 $t+1$ から時点 t へのクラスタ一致率によるクラスタの変化、データラベルの同一性に基づいたクラスタ一致率を提案する。そして時系列データが容易に取得できるMLBデータを用いて、提案手法がクラスタリング手法に依存しないことを確認するため複数の手法で実験を行う。

2. 関連研究

2.1 動的クラスタリング

クラスタリングの対象となるデータ集合が時系列ごとにデータ数の変化が起こることがあり、時系列ごとにクラスタリングを行う必要がある。そのようなクラスタリングを動的クラスタリングと呼ぶ。動的クラスタリングではデータ数の変化により、所属クラスタの変化、クラスタ数の増

¹ 名古屋工業大学
Nagoya Institute of Technology

² 中部大学
Chubu University

a) k.yoneda.690@nitech.jp

減が時系列ごとにかかることがある。クラスタの特徴を把握するため、クラスタの変化を見る研究 [1], クラスタリング構造変化点検出に関する研究 [2] など動的クラスタリングに関する研究 [3], [4], [5] は数多くある。

Spiliopoulou ら [1] は動的クラスタリングの時系列で起こりうるクラスタの変化を集合の変化と捉え、クラスタリング全体の構造に着目せずに、クラスタを構成する要素に着目したクラスタの変化に関する研究を行った。水野ら [3] らの過去のクラスタリング結果を現在のデータへ割り当てる制約付きクラスタリングとは異なり、過去の結果を考慮せずに動的クラスタリングを行うため、クラスタリング手法に依存する必要がない。

2.2 MONIC

Spiliopoulou ら [1] は連続的にデータが増える環境において、1回のクラスタリングを行うにあたり、データエイジング関数により期間を指定し、指定した期間のデータを混ぜてクラスタリングを行い、データの同一性に基いて定義したクラスタの一致率を用いて、時系列ごとにクラスタの変化を分析した。MONIC ではデータの同一性に基いて「クラスタ内データの特徴の類似性」をクラスタの変化として見る事が出来た。x を実験対象となるデータとしたとき、時点 t から見た x のデータエイジング関数は以下となる。

$$\text{age}(x, t) \in [0, 1]$$

時点 t, t' ($t < t'$) のクラスタの変化を見るときに時点 t のクラスタから時点 t' のクラスタへのデータの同一性を用いたクラスター一致率 overlap を定義した。そして、時点 t のクラスタ $C_{t,a} \in \Gamma_t$ から見たクラスター一致率で最大となる時点 t' のクラスタ $C_{t',b} \in \Gamma_{t'}$ が存在し、閾値 τ を超えた場合、その2クラスタはマッチしているとし、 $C_{t',b} = \text{match}_\tau(C_{t,a}, \Gamma_{t'})$ と表した。ここで、 Γ_t は時刻 t におけるクラスタの集合を表す。クラスタ $C_{t,a} \in \Gamma_t$ と $C_{t',b} \in \Gamma_{t'}$ のクラスター一致率を以下に示す。

$$\text{overlap}(C_{t,a}, C_{t',b}) = \frac{\sum_{e \in C_{t,a} \cap C_{t',b}} \text{age}(e, t')}{\sum_{d \in C_{t,a}} \text{age}(d, t')}$$

MONIC ではクラスター一致率、マッチを用いてクラスタの変化として生存、分裂、吸収、消滅の4種類の変化を定義した。生存で用いる閾値 τ と分裂で用いる閾値 τ_{split} は $\tau_{\text{split}} < \tau \in [0.5, 1]$ とする。クラスタの変化を表 1 に示す。

2.3 MONIC の改善すべき点

本研究では時系列データを取得するために、MLB のチームごとの打撃・守備データに着目した。MLB データでクラスタリングした場合、打撃が強いチームが集まるクラスタは次時点で打撃が強いチームのクラスタと一致率が高くなる。しかし、打撃が強いクラスタが翌年に向けて守備力

を強化し、打撃と守備力が強いクラスタへ変化した場合、一致率を示すことが出来ず、クラスタの変化を表せれない。また、MLB では年代ごとに野球の流行りが存在するため、クラスタリングを行うと各年代でクラスタが形成されることがある [7]。このような場合、一致率を示すことが出来ず、クラスタの変化を表せれない。加えて、クラスター一致率は前の時点からのクラスタの変化しか見ていない。これにより正確なクラスタの変化を見る事が出来ない。表 2 は Γ_1 と Γ_2 のクラスター一致率を表している。行は時点 2、列は時点 1 のクラスタを表している。表 2 から $C_{1,1}$ と $C_{2,1}$ の一致率は 0.6 であり、 $C_{1,1}$ に含まれるデータのうち、6割が $C_{2,1}$ に含まれていることを示している。 $C_{1,1}$ は $C_{2,1}$ とマッチしており、 $C_{2,1}$ は他に時点 1 のクラスタとマッチしていないので、 $C_{1,1}$ は $C_{2,1}$ として生存している。ここで、 $C_{1,1}$ の要素数が 10 で $C_{1,2}$ が 100 の場合、 $C_{2,1}$ の要素のうち、生存している $C_{1,1}$ からは 6 個、生存していない $C_{1,2}$ からは 30 個となる。MONIC による分析では $C_{2,1}$ は生存していない $C_{1,2}$ の要素が多く含まれているにもかかわらず、 $C_{1,1}$ と生存を示している。MONIC ではこのようなクラスタの変化を十分に表すことが出来ない。

3. 提案手法

3.1 提案手法のアプローチ

本研究では「クラスタ内データの特徴の変化の類似性」をクラスタの変化として見ることを目的に、データラベルの同一性に基いたクラスター一致率を提案する。打撃が強いクラスタから打撃・守備が強いクラスタへクラスタごと移動した場合のクラスタ変化分析が可能になる。加えて、時点ごとにクラスタリングを行う動的クラスタリングを行うことで、MONIC で起こりうる各年代でクラスタを形成する事例がなくなる。また、表 2 で説明したクラスタ変化の問題を変化後のクラスタを基準としてクラスター一致率を見ること（後ろ向き分析）で解決する。

3.2 データモデルとクラスター一致率

本研究では、1時点でデータラベルの重複がなく、時点ごとにラベルの増減が少ないデータに対する問題設定のもと、クラスタの変化を見る。本研究で扱うデータモデルを図 3、データラベルの同一性に基いたクラスター一致率を以下に示す。生存で用いる閾値 τ と分裂で用いる閾値 τ_{split} は $\tau_{\text{split}} < \tau \in [0.5, 1]$ とする。

$$\text{overlap}(C_{t,a}, C_{t',b}) = \frac{|L_{t,a} \cap L_{t',b}|}{|L_{t,a}|}$$

3.3 クラスタの変化

本研究では、時点 t から t' ($t < t'$) へのクラスタの変化 4 種類（前向き分析）に、加えて時点 t' から分析したクラスタの変化を 2 種類（後ろ向き分析）見ることで複雑なクラ

表 1 MONIC でのクラスタの変化
Table 1 Cluster change in MONIC

$C_{t,a}$ が $C_{t,b}$ として生存	$C_{t,b} = \text{match}_\tau(C_{t,a}, \Gamma_{t'}) \text{ AND } \nexists C_{t,c} \in \Gamma_t \setminus \{C_{t,a}\} [C_{t,b} = \text{match}_\tau(C_{t,c}, \Gamma_{t'})]$
$C_{t,a}$ が $\{C_{t',1}, \dots, C_{t',p}\}$ に分裂	$\emptyset = \text{match}_\tau(C_{t,a}, \Gamma_t) \text{ AND } (\forall u = 1, \dots, p \text{ overlap}(C_{t,a}, C_{t',u}) \geq \tau_{\text{split}}) \text{ AND } \text{overlap}(C_{t,a}, \cup_{u=1}^p C_{t',u}) \geq \tau \text{ AND } (\nexists C_{t',c} \in \Gamma_t \setminus \{C_{t',1}, \dots, C_{t',p}\} [\text{overlap}(C_{t,a}, C_{t',c}) \geq \tau_{\text{split}}])$
$C_{t,a}, C_{t,c}$ が吸収されて $C_{t,b}$ になる	$C_{t,b} = \text{match}_\tau(C_{t,a}, \Gamma_{t'}) \text{ AND } \exists C_{t,c} \in \Gamma_t \setminus \{C_{t,a}\} [C_{t,b} = \text{match}_\tau(C_{t,c}, \Gamma_{t'})]$
$C_{t,a}$ は消滅	上記のクラスタの変化に当てはまらない場合

表 2 Γ_1 と Γ_2 のクラスター一致率

Table 2 Cluster matching rate of Γ_1 and Γ_2

	$C_{2,1}$	$C_{2,2}$	$C_{2,3}$	$C_{2,4}$	$C_{2,5}$
$C_{1,1}$	0.6	0.2	0.1	0	0.1
$C_{1,2}$	0.3	0.7	0	0	0

表 3 本研究におけるデータモデル

Table 3 Data model in this study

時点 t のデータ集合 $D_t = \{d_{t,1}, d_{t,2}, \dots, d_{t,n}\}$
 時点 t のクラスタ集合 $\Gamma_t = \{C_{t,1}, C_{t,2}, \dots, C_{t,k}\}$
 $l_{t,a}$ データ $d_{t,a}$ のデータラベル
 $L_{t,b}$ クラスタ $C_{t,b}$ に含まれるデータラベル集合

表 4 クラスタ変化の後ろ向き分析

Table 4 Backward analysis of cluster change

$C_{t',b}$ は $C_{t,a}$ の影響が強い	$C_{t,a} = \text{match}_\tau(C_{t',b}, \Gamma_t)$
$C_{t',b}$ は $C_{t,a}, \dots, C_{t,p}$ で構成される	$\emptyset = \text{match}_\tau(C_{t',b}, \Gamma_t) \text{ AND } (\forall u = 1, \dots, p \text{ overlap}(C_{t',b}, C_{t,u}) \geq \tau_{\text{split}}) \text{ AND } \text{overlap}(C_{t',b}, \cup_{u=1}^p C_{t,u}) \geq \tau \text{ AND } (\nexists C_{t',c} \in \Gamma_t \setminus \{C_{t,1}, \dots, C_{t,p}\} [\text{overlap}(C_{t',b}, C_{t,c}) \geq \tau_{\text{split}}])$

スタの変化を明確にする。

表 4 に後ろ向き分析を用いたクラスタの変化を示す。前後の分析を行うことで以下のクラスタの変化が分かる。表 2 は Γ_1 と Γ_2 のクラスター一致率を表している。行は時点 1, 列は時点 2 のクラスタを表している。表 4 から $C_{2,1}$ と $C_{1,2}$ の一致率は 0.7 であり, $C_{2,1}$ に含まれるデータのうち, 7 割が $C_{1,2}$ にも含まれていることを示している。 $C_{2,1}$ は $C_{1,2}$ とマッチしているため, $C_{2,1}$ は $C_{1,2}$ の影響が強いとなる。そして, 表 2 での前向き分析で $C_{1,1}$ は $C_{2,1}$ として生存している。前後の分析から, $C_{2,1}$ は生存している $C_{1,1}$ よりも $C_{1,2}$ の影響を受けてクラスタを形成していることが示唆される。このように, 前後の分析を行うことで前向きでは対応出来ないクラスタ変化を見ることが可能になる。

4. 実験と評価

本章では実験に用いるデータの説明を行った後, t-sne を

表 5 Γ_1 と Γ_2 のクラスター一致率 (後ろ向き) の例

Table 5 Example of cluster matching rate of Γ_1 and Γ_2 (Backward)

	$C_{1,1}$	$C_{1,2}$
$C_{2,1}$	0.3	0.7
$C_{2,2}$	0.55	0.45

用いてデータの次元数を 2 次元まで減らし, クラスタリングの安定性を証明した後にクラスタスタリングを行い, 結果をクラスタごとに色分けしたネットワーク図で示す。そして, クラスター一致率を用いてクラスタの変化を分析する。従来手法では 2 年ごとにデータを混ぜて MLB チームのクラスタリングを行い, 提案手法では単年のデータで MLB チームのクラスタリングを行う。提案したクラスター一致率を用いたクラスタの変化 (前向き分析) と Spiliopoulou ら [1] が提案したクラスター一致率を用いたクラスタの変化 (MONIC) の比較を行う。さらに, 前向き分析で見ることが出来ないクラスタの変化を後ろ向き分析で補えることを示す。

4.1 実験データ

SeanLahman.com にある Lahman の Baseball Database (<http://seanlahman.com/baseball-archive/statistics>) の 2007 年から 2016 年までのシーズンごとの球団の成績を含む Teams データを使用した。その中で, どのデータかを示す yearID, チーム名を示す teamID, 得点数, 盗塁数, 三振数など打撃データ, 防御率, 奪三振数, 被安打数など守備データを使用した。勝利数や優勝, ワイルドシリーズ進出など直接勝敗や順位が分かるデータは使用していない。yearID は時刻として用い, teamID はデータのラベルとして用いた。打撃データ, 守備データは属性として用いた。

まず, 全てのチームの全ての年に対して, 打撃データと守備データを tSNE を用いて 2 次元に変換した。次に, 2 次元に変換したデータに対して, 年ごとにクラスタリングを行い, 動的クラスタリングを行った。

4.2 分割数

実験で用いるクラスタリング手法は k-means 法と, x-means 法 [9] である。そのうち, k-means 法は予めクラスタ分割数を決める必要があり, 適切なクラスタ分割数を事前に把握し決める必要がある。本研究では野球のチーム性

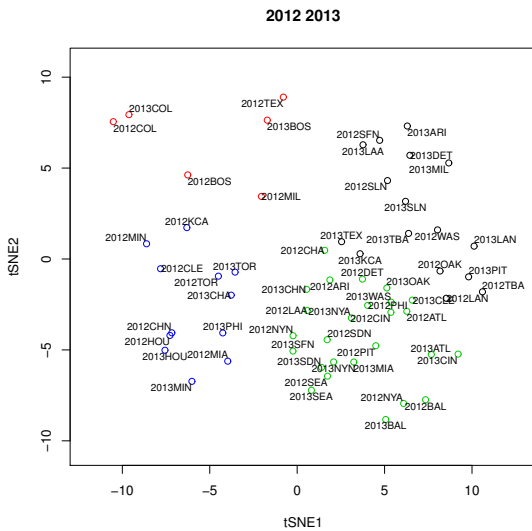


図1 k -means 法でのクラスタリング (2012-2013 年)
Fig. 1 Clustering by k -means method (2012-2013)

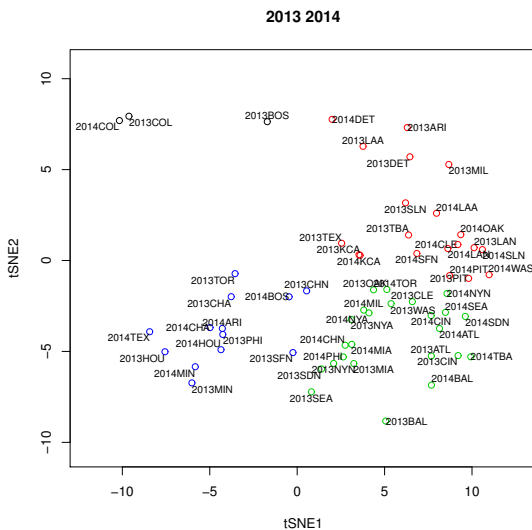


図2 k -means 法でのクラスタリング (2013-2014 年)
Fig. 2 Clustering by k -means method (2013-2014)

質が「打撃力、守備力がともに高いチーム」、「打撃力が高く、守備力の低いチーム」、「守備力が高く、打撃力が低いチーム」、「打撃力、守備力がともに低いチーム」の大きく分けて4つあると考え、クラスタ分割数を4とした。

4.3 実験結果

4.3.1 MONIC でのクラスタリング

MONIC では連続的にデータが取得できる環境において t 年から $t+1$ 年までの2年分データごとにまとめて k -means 法でクラスタリングを行った。そのため、今回の実験でも MLB データ 2007 年から 2016 年のデータを {2007,2008},{2008,2009},...,{2015,2016} ごとにデータをまとめて k -means 法でクラスタリングを行う。図1, 2は {2012,2013},{2013,2014} のデータを k -means 法でクラス

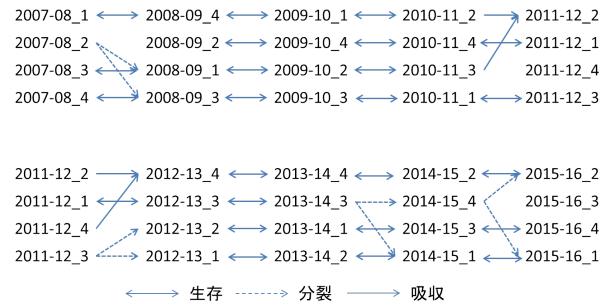


図3 MONIC によるクラスタの変化 (前向き分析)
Fig. 3 Cluster change by MONIC(Forward analysis)

タリングした結果です。ノードラベルはチーム名とデータの取得年を表している。横軸が第一成分、縦軸が第二成分を表しており、各軸の長さはすべての結果で統一させているため、データの違うクラスタリング結果でもデータの位置が似ていれば類似していることが図を見て分かることが出来る。各結果のクラスタの色は結果間で対応付けていない。図12を見ると2014年のデータが図の右側に偏りがあることが分かる。このことから、年ごとに打者有利や投手有利などの傾向があることが示唆される。

4.3.2 MONIC でのクラスタの変化

図3は、クラスタの変化を時系列ごとに並べたものである。クラスタの変化を示すときに用いた閾値 $\tau = 0.6$, $\tau_{split} = 0.2$ とした。両方に矢印がついているクラスタ間は「生存」していることを示し、点線で右矢印がついているクラスタ間は「分裂」の関係を示し、普通の線で右矢印がついているクラスタ間は「吸収」の関係を示す。

4.3.3 k -means 法を用いた提案手法

k -means 法でのクラスタリング結果が図4, 5である。図4は2008年のMLBデータを、図5は2009年のMLBデータを k -means 法でクラスタリングした結果です。ノードラベルはチーム名を表している。横軸が第一成分、縦軸が第二成分を表しており、各軸の長さはすべての結果で統一させているため、データの違うクラスタリング結果でもデータの位置が似ていれば類似していることが図を見て分かることが出来る。各結果のクラスタの色は結果間で対応付けていない。

4.3.4 提案手法でのクラスタの変化

時刻 t から $t+1$ への前向き分析によるクラスタの変化を図6、後ろ向き分析を図7に示す。クラスタの変化を示すときに用いた閾値 $\tau = 0.6$, $\tau_{split} = 0.2$ とした。クラスタ名は2007.1のように、「データ年数_クラスタ番号」とした。

4.3.5 x -means 法を用いた提案手法でのクラスタの変化

時刻 t から $t+1$ への前向き分析によるクラスタの変化を図8、後ろ向き分析によるクラスタの変化を図9に示す。 x -means 法 [9] はデータに対して適切とされるクラスタ数で k -means 法を行う手法のため、クラスタ数にばらつき

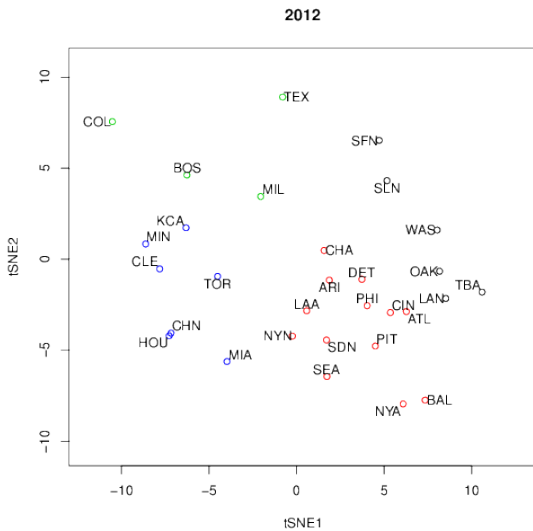


図4 k-means 法でのクラスタリング (2012 年)
Fig. 4 Clustering by k-means method (2012)

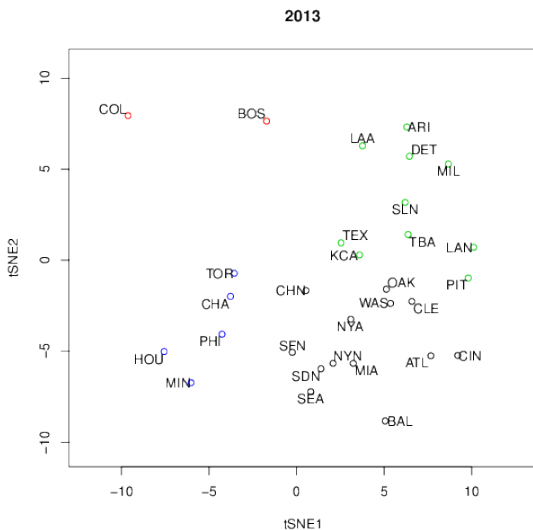


図5 k-means 法でのクラスタリング (2013 年)
Fig. 5 Clustering by k-means method (2013)

がある。2013 年と 2014 年を見るとクラスタ分割数は 3 から 2 に減っている。そして、2013.1 と 2013.2 が吸収されて 2014.1 となっていることが分かった。このことからクラスタリング結果ごとにクラスタ数が増えた環境でもクラスタの変化を見ることが確認できた。2015 年から 2014 年のクラスタの変化を見ると 2015.4, 2015.3, 2015.2 が 2014.1 から「影響」を受けていることがわかった。クラスタ内要素数が多いクラスタの変化を見るとき、クラスタ内要素数の大きいクラスタから「生存」しているクラスタの要素数も大きいため、他のクラスタ占める割合が大きいのに関わらず、クラスタの変化として見ることが出来ないことが今回の実験で確認できた。この問題には後ろ向きのクラスタの変化を見ることで解決出来ることも確認できた。

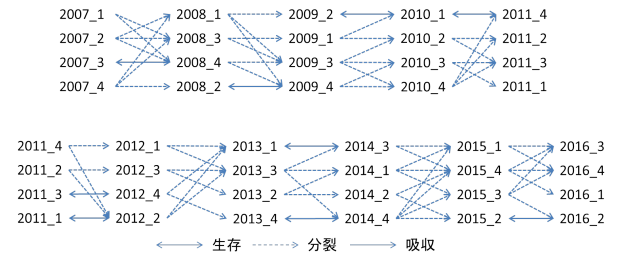


図6 前向き分析によるクラスタ変化
Fig. 6 Cluster changes by forward analysis

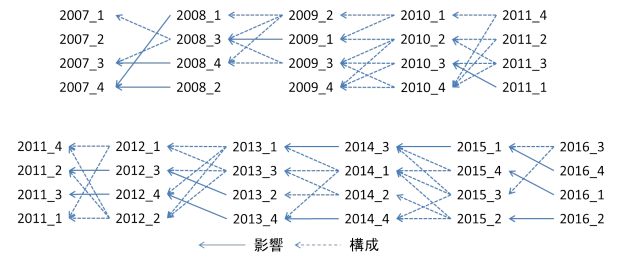


図7 後ろ向き分析によるクラスタ変化
Fig. 7 Cluster changes by backward analysis

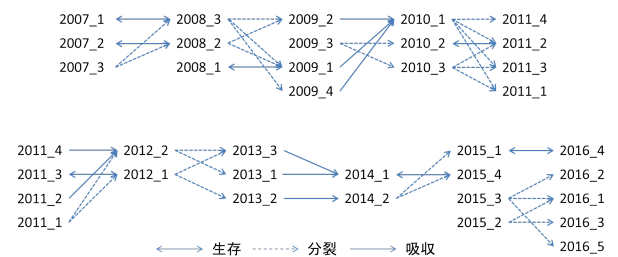


図8 x-means 法を用いた前向き分析によるクラスタ変化
Fig. 8 Cluster change by forward analysis using x-means method

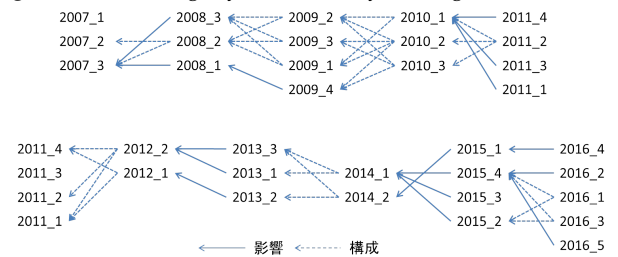


図9 x-means 法を用いた後ろ向き分析によるクラスタ変化
Fig. 9 Cluster changes by backward analysis using x-means method

5. 考察

5.1 MONIC と前向き分析の比較

MONIC では 2012-13.2 と 2013-14.1 が生存しているが、図1, 図2を見ると 2012-13.2 (赤クラスタ) には 2012 年の BOS, COL, MIL, TEX と 2013 年 BOS, COL が所属している。そして 2012-13.1 (黒クラスタ) には 2013 年の MIL, TEX が含まれていることから、MIL, TEX は 2012 年から 2013 年にかけてチーム方針を変えており、BOS, COL はチーム方針を変えていないことが分かる。そして、

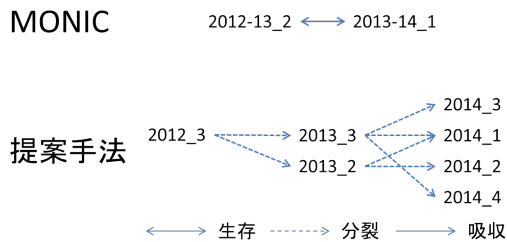


図 10 MONIC と提案手法のクラスター変化の比較

Fig. 10 Comparison of cluster change of MONIC and proposed method

2013-14 では 4 チーム異なるクラスターに所属していることから 2012 年から 2014 年にかけてチーム方針が大きく変化していることが分かる。しかし、MONIC では説明したクラスター変化を見ることが出来ない。一方、提案手法では BOS, COL, MIL, TEX が所属する 2012.3 から BOS, COL が所属する 2013.2 と MIL, TEX が所属する 2013.3 に分裂した。さらに 2013.2, 3 は 2014 年のクラスターすべてに分裂した。このことから、従来手法では見る出来ないクラスター変化を提案手法では見ることが出来た。

5.2 前向き分析と後ろ向き分析の比較

図 7 は後ろ向き分析で見たクラスター変化を示している。閾値など条件は前向き分析と同じである。図 6, 図 7 を見ると 2007.3 は 2008.4 として「生存」しており、2008.4 は 2007.3 からの「影響」を受けていることが分かる。これにより、クラスター 2007.3 はクラスター内要素数に大きな変化がない状態で 2008.4 として完全に「生存」していることが分かる。2007.2 は 2008.1, 2008.3, 2008.4 に「分裂」したが、逆から見ると 2008 年のクラスターのどれともクラスター一致率が 0.2 を超えるものはいなかったことが分かる。このことから 2007.2 はクラスター内要素数が少ないクラスターであるため、2008 年の 3 つクラスターにも少ない数しか入っていないことが分かる。2009.2 を見ると、2010.1 として「生存」していることが分かる。また、2009.1 は 2010.1 と 2010.2 に分裂したことが分かる。2010.1 は「生存」している 2009.2 の「影響」があると思いきや、逆から見ると、2009.1 の「分裂」された要素と 2009.2 で構成されていることが分かった。原因として、2009.2 のクラスター内要素数が少ないため、「生存」と判定されやすいからである。このように、前向き分析のクラスターの変化だけではわからないクラスターの変化も後ろ向き分析のクラスターの変化を加えることで見つけることが出来る。

5.3 クラスタリング手法による提案手法の比較

本研究で用いたクラスタリング手法は k -means 法, x -means 法である。 k -means 法で実験した時のクラスターの変化を見た際、分割が多いことが見られた。原因として、

データに対して適切な分割数が固定した数字と違うためだと考えられる。指定した分割数が適切な数より多い場合、無理やり多くのクラスターに分割するため、本来であれば同じクラスターに含まれているべきデータがそれぞれ別のクラスターに含まれてしまう。それにより、クラスター一致率が低くなりクラスターの変化が「分裂」となりやすい傾向にあると考えられる。このことから、適切なクラスター分割数でクラスタリングを行う x -means 法が本研究に適したクラスタリング手法であると示唆される。

6. まとめと今後の課題

本研究では「クラスター内データの特徴の変化の類似性」をクラスターの変化として見ることを目的に、後ろ向きのクラスター一致率によるクラスターの変化の検出と、データラベルの同一性に基づいたクラスター一致率を提案した。MONIC では見ることが出来なかったクラスターの変化を確認した。

今後の課題として、クラスターの変化で用いる閾値の適切な値を調査すること、任意の時点間のクラスターの変化を見ることで前後以外のクラスターの関係性が発見出来ると考えている。

参考文献

- [1] Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., et al, MONIC-Modeling and Monitoring Cluster Transitions KDD'06, pp. 20-23 (2006)
- [2] So Hirai, Kenji Yamanishi, Detecting Changes of Clustering Structures Using Normalized Maximum Likelihood Coding, KDD'12, pp. 12-16 (2012)
- [3] 水野珠季, 廣安知之, 三木光範ほか, 制約付きクラスタリングによるデータの時系列変化の把握, 第 23 回人工知能学会全国大会, 2B3-3 (2009)
- [4] 天野恵理子, 大枝真一, クラスタ数の変化点検出を用いた授業アンケートの分析, 情報処理学会第 75 回全国大会, No. 4, pp. 689-690 (2013)
- [5] 福井 健一, 斉藤 和巳, 木村 昌弘ほか, 自己組織化ネットワークによる動的クラスタの可視化編纂, 人工知能学会論文誌, 23 巻, 5 号, pp. 319-329 (2008)
- [6] Keke Chen, Ling Liu, HE-Tree: a framework for detecting changes in clustering structure for categorical data streams, The VLDB Journal, 18:1241-1260 (2009)
- [7] Melling, M., Clustering (Almost) Every MLB Team... Ever, Baseball Data Science, <http://www.baseballdatascience.com/clustering-almost-every-mlb-team-ever/> (2016)
- [8] Schmidt, M. B., Berriz D. J., Convergence and clustering in Major League Baseball: the haves and have nots?, Applied Economics, 36(18):pp2007-2014 (2004)
- [9] 石岡恒憲, クラスタ数を自動決定する k -means アルゴリズムの拡張について, 応用統計学, Vol. 29, No. 3, pp. 141-149 (2000)
- [10] 鈴木 努, 「R で学ぶデータサイエンス 8 ネットワーク分析」, 共立出版, (2009)