ARG WI2 No.14, 2019
WI2-2019-12

多次元時系列データに対する評価指標形成のための視覚的分析フレームワークの提案

高見 玲[†,a]　　高間 康史[†,b]

† 首都大学東京 システムデザイン研究科

a) *takami-rei@ed.tmu.ac.jp*　b) *ytakama@tmu.ac.jp*

概要　本稿では, 多次元時系列データに対する指標形成の支援を目的とした視覚的分析フレームワークを提案する．評価指標は，データに基づく意思決定，機械学習のためのラベル付けなどにおいて重要な役割を果たす．指標の形成はドメイン専門家の視覚的分析を通して行われるのが一般的だが, 多次元時系列データの場合, 分析者が次元削減等のアルゴリズムやパラメータを理解することが困難である．そのため, 可視化技術と専門家の知識を融合するための, 相互主導型インタフェースの導入が期待される．提案するフレームワークでは，属性値の重み付き線形和の形で, 時点ごとにデータの評価指標を相互主導型インタラクションにより形成する．提案フレームワークに基づきインタフェースを実装し，実データに適用した例を示す．

キーワード　時系列データ, 視覚的分析, 相互主導型インタフェース, 情報可視化インタフェース, 評価指標

1　はじめに

多次元時系列データは, Web サービスのログデータや医療, スポーツなど, 多様な分野で収集・分析対象として普及している．これらの意思決定や機械学習への活用には, 外れ値の決定において重視する属性値を決定するための評価基準や, クラスタ間距離のような評価指標が重要となる．評価指標は, 各データドメインの専門家が対象データに関する知識に基づき決定する．データが有する特性を専門家が理解する必要のある指標形成を支援するため, 視覚的分析インタフェースの有効性が期待される．しかし，指標形成支援を明示的な目的としたインタフェースの研究はこれまで行われていない．

本稿では, 多次元時系列データの評価指標形成を, 直接操作により支援する視覚的分析フレームワークを提案する．同一目的でも時点毎に異なる指標が形成されうることを想定し, アニメーションで時系列性を表現した散布図への直接操作に基づき, 時点毎に異なる評価指標を形成する．指標の形態として, 属性値の重み付き線形和を想定する．この形態は解釈容易性に優れ, 主成分分析などの次元削減手法との親和性が高いといった利点がある．そして, 提案フレームワークを視覚的分析インタフェースとして実装し, 実データを用いた分析事例を示す．

2　関連研究

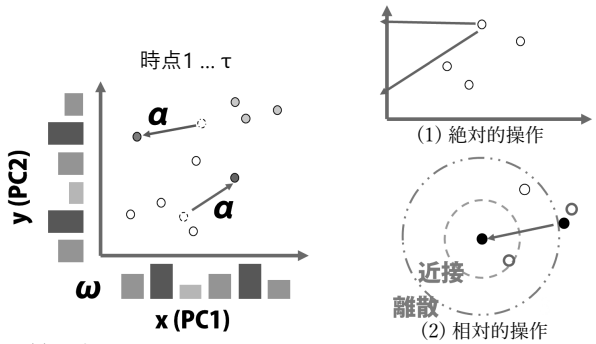
2.1　視覚的分析

ドメイン専門家は分析対象データに関する知識を持つが, 必ずしもデータ分析に詳しくはない．そのため, システム側がデータを可視化することで, ユーザの視覚的な認知能力を活用した洞察形成を支援する視覚的分析インタフェース [1] が用いられる．しかし, 前処理として用いられるクラスタリングアルゴリズムや, 次元削減手法等の背後に存在するモデルやパラメータ調整の困難性が指摘されている [2]．この問題に対処するため, システムの計算処理能力とユーザの認知能力を組み合わせ, 対象の作業を効率的に行う相互主導型インタフェースが提案されている [3]．中でも, 可視化オブジェクトへの直接操作の意図をシステム側が解釈してアルゴリズムのパラメータ調整を行い, 再処理結果をユーザにフィードバックする Semantic Interaction のコンセプトが提案されている [4]．これによって, ユーザは自身のドメイン知識を可視化に反映する形式でパラメータ調整を行える．

本稿で扱う多次元時系列データ特有の問題として, インタラクションの複雑化に伴うユーザの混乱や, 可視化手法間のトレードオフが挙げられる．これらの問題に対し, 時系列データの軌跡表現による視覚的傾向の可視化や, 凸包表現によるデータのグルーピングを用いて視覚的分析を支援するインタフェースが提案されている [5]．しかし, 多次元データへの対応や, Semantic Interaction によるドメイン知識の反映は考慮されていない．

2.2　評価指標形成

本稿では, ユーザが定義した基準に対して, 対象データがどの程度“重要”かを各属性値の組成で表現したものを評価指標と定義する．視覚的分析を通じて指標形成を行うためには, ユーザは可視化に影響するパラメータを調整しながら, 基準との一致度を主観的に判断し, 漸進的に指標を構築する必要がある．多次元データに対するパラメータ調整の支援例として, Kim らは, 指定したオブジェクトが強調されるように, 線形次元削減結果の各属性の重みを調整するインタフェース InterAxis を提



(a) 提案フレームワークのパラメータ　(b) ユーザ操作の解釈手法

図 1 提案インタフェースの分析パラメータ

案している [6]. Wall らは, ランキング結果のテーブルを並び替える操作に基づき, ランキングの再学習を行うインタフェース Podium を提案している [7]. しかし, 時系列データは可視化する際に時間的変化の表現手法などの多くの考慮すべき特性を持つ [8] 上, その評価指標はデータの周期性や時点による傾向の違いを考慮するべきである. そのため, 上述の既存研究をそのまま時系列データに対し適用することは困難と考える.

指標形成を明示的に支援するための取り組みとして, Web サービスのユーザ行動に対する評価指標構築が研究されている [9]. この研究では, 特定の操作を行ったユーザのプロフィールなどを含む多くの特徴量から, 分類に有効なものを機械学習により抽出し, 評価指標の構成要素の候補として提示することで指標形成を支援している. しかし, 最終的な指標に用いる特徴量の選定や, それらの組成による指標形成は専門家により行われるため, 指標形成のための探索的分析を支援するものではない.

3 提案フレームワーク

フレームワークの適用対象として, 既存かつ有限の多次元時系列データを想定する. 対象データ $d_{\tau nm}$ は N ($n \in \{1,2,...,N\}$) 個の M ($m \in \{1,2,...,M\}$) 次元ベクトルで構成され, 時点 $\tau \in \{1,2,...,T\}$ 毎に離散的に記録される. このデータを, τ 毎に 2 次元に次元削減した上で散布図として可視化する. ユーザは散布図上のオブジェクトを直接操作し, システムは操作内容に基づき投影軸における各属性の寄与度合いを調整し, 散布図を再描画する. また, 各属性の寄与度合いを指標やその叩き台として明示的にフィードバックする. この手順を繰り返し, ユーザは漸進的に指標を構築できる.

3.1 分析パラメータ

図 1 に, 提案するフレームワークの分析パラメータと, それらを用いた Semantic Interaction の概観を示す. 散布図上での時系列データ $\boldsymbol{d_{\tau n}} \in \mathbb{R}^M$ の 2 次元座標 $\boldsymbol{P_{\tau n}} = (X_{\tau n}, Y_{\tau n})$ は, 式 (1)(2) のように, パラメータ $\boldsymbol{\alpha}, \boldsymbol{\omega}$ により定義される (図 1(a)). $\boldsymbol{\alpha}, \boldsymbol{\omega}$ は散布図の 2 軸について変更可能なため, それらを上付き文字 X, Y のいずれかで表している. $\boldsymbol{d_{\tau n}}$ の X, Y 軸に対するパラメータはそれぞれ, $\alpha^{\mathrm{X}}_{\tau n}, \alpha^{\mathrm{Y}}_{\tau n}$ で表現される. $\boldsymbol{\omega}$ は全時系列データに共通であり, m 番目の属性に対応する X, Y 軸におけるパラメータは $\omega^{\mathrm{X}}_{\tau m}, \omega^{\mathrm{Y}}_{\tau m}$ で表される.

- $\boldsymbol{\alpha} \in \mathbb{R}^{2\times N\times T}$: $\boldsymbol{d_{\tau n}}$ の散布図上における強調度合いを表現する移動度 (バイアス).
- $\boldsymbol{\omega} \in \mathbb{R}^{2\times M\times T}$: 投影軸への寄与度を定義する各属性の係数.

$$X_{\tau n} = \sum_{m=1}^{M} d_{\tau nm}\omega^{\mathrm{X}}_{\tau m} + \alpha^{\mathrm{X}}_{\tau n} \quad (1)$$

$$Y_{\tau n} = \sum_{m=1}^{M} d_{\tau nm}\omega^{\mathrm{Y}}_{\tau m} + \alpha^{\mathrm{Y}}_{\tau n} \quad (2)$$

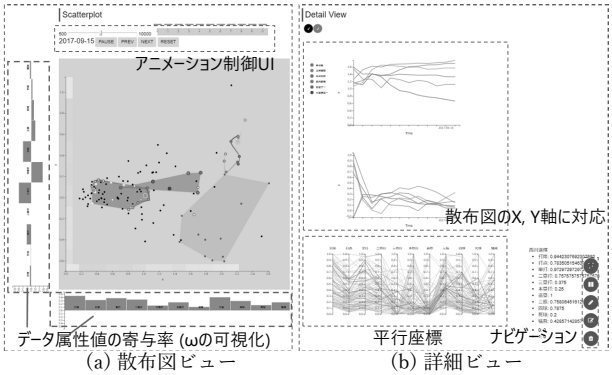
3.2 評価指標

提案フレームワークにおける評価指標はパラメータ $\boldsymbol{\omega}$ に対応する. 属性値の線形結合による指標は, データの非線形性を表現できない問題を有するが, 本稿では結果の解釈可能性や計算コストを優先して線形結合を採用する. また, 野球におけるセイバーメトリクス等における使用例から, 線形和に基づく指標の形成は十分な実用性があると判断する. 時系列データの評価指標は, 季節などの周期性や, 特定時点での傾向の違いを考慮する必要があると考え, 時点毎に $\boldsymbol{\omega}_t$ を修正可能にする.

3.3 指標の変更

Bernerd らは, オブジェクト間の類似度の変更対象を可視化画面における絶対的位置 (absolute), オブジェクト周囲の制約半径 (orbital), オブジェクト間の相対的な位置関係 (relative) の 3 つに分類している [10]. 提案フレームワークでは, タスクへの妥当性や, 単純化のため, $\boldsymbol{\alpha}$ と $\boldsymbol{\omega}$ の変更に対応する以下の 2 つの操作を散布図へのユーザ意図の表現手法として想定する (図 1(b)).

- 絶対的操作: 散布図の X, Y 軸に対する絶対的な位置指定に基づく, 投影軸の大局的な変更を行う. ユーザが散布図の X, Y 軸へ対象オブジェクトをドラッグ&ドロップすると, システムは当該オブジェクトが投影軸上のドロップ位置付近に配置されるように各属性の $\boldsymbol{\omega}$ を調整し, 調整結果をユーザにフィードバックする.
- 相対的操作: 散布図内での相対的な位置関係の変更による, 投影結果の局所的な調整を行う. ユーザが散布図内の任意の位置にオブジェクトをドラッ



(a) 散布図ビュー (b) 詳細ビュー

図 2 提案インタフェースの画面構成

(a) 軌跡表現 (b) 凸包 (アニメーション再生時) (c) 凸包 (アニメーション静止時)

図 3 操作対象オブジェクト

グ&ドロップすると, システムは当該オブジェクトがドロップ位置に配置されるように $\boldsymbol{\alpha}$ を修正する. $\boldsymbol{\alpha}$ は個々のデータへのバイアスであり, 指標に反映するには最適化計算などを用いて $\boldsymbol{\omega}$ に還元する必要がある.

両操作を繰り返し, ドメイン知識に合致する散布図が得られたときの $\boldsymbol{\omega}$ を評価指標, あるいはその叩き台として利用する.

4 提案インタフェース

提案フレームワークの Web アプリケーションとしての実装例を図 2 に示す. アプリケーション全体は Ruby on Rails 5 を, 可視化とフロントエンド部分は D3.js v5 と jQuery 3.3 を用いて実装し, パラメータ調整は Python 3 で実装されたコードを Flask[1] で呼び出して行う. 結果が属性値の重み付き線形結和で表される特性や, 計算コストの低さを重視して, 次元削減手法には主成分分析を採用した. 提案インタフェースは, 散布図ビューと詳細ビューから構成される.

- 散布図ビュー (図 2(a)): 多次元時系列データの指定時点における次元削減結果を, アニメーション制御可能な 2 次元散布図で可視化する. データの時系列的傾向に関する洞察を得るために用いる.
- 詳細ビュー (図 2(b)): 散布図ビュー上で選択されたオブジェクトの属性値やメタデータを線グラフや並行座標で表示する. 散布図で得られた視覚的洞察の検証に基づく仮説形成を行うために用いる.

4.1 探索モード

提案フレームワークではアニメーションの再生/静止を探索モードとして導入する. 再生モードでは, 各時点における時系列データの散布図を一定間隔でアニメーションし, 静止モードでは特定の τ に散布図を固定する. ユーザは再生モードにおける散布図のアニメーションから得られる視覚的概要に基づき, 静止モードにて各オブジェクトの詳細な探索を行い, 両モード間の遷移を活用して分析を行うと想定する. ユーザインタラクションの解釈を探索モードにより分岐させることで, 各モードにおける分析行動に適した操作を実現する.

4.2 操作対象オブジェクト

ドラッグ&ドロップやクリックなどの直接操作対象となるオブジェクトは, 以下の 3 種類に分類される.

- データ点: データ $\boldsymbol{d}_{\tau n}$ の散布図上の点 $\boldsymbol{P_{\tau n}}$ を表す. データ点に 3.3 節のような直接操作を行うと, パラメータは対応する τ についてのみ調整される.
- 軌跡: 図 3(a) に示すような, $P_{1n}, P_{2n}, ..., P_{Tn}$ を曲線で補完した表現であり, 散布図上の点のマウスオーバーで表示される. 軌跡の各ノードをクリックすると, 再生時点に対応する時点へ散布図が更新される. 軌跡に対して直接操作を行うと, パラメータは軌跡が含む全ての τ について調整される.
- 凸包: オブジェクトの集合を表す. 再生モードでは図 3(b) のような再生時点での複数データの凸包が, 静止モードでは図 3(c) のような複数データ点の軌跡の凸包が描画される. 凸包は散布図上のデータ点を囲う曲線の描画により作成され, ユーザはその形状から複数オブジェクト間の関係を把握できる. 凸包の直接操作に基づくパラメータ調整は, その内部の全オブジェクトに適用される.

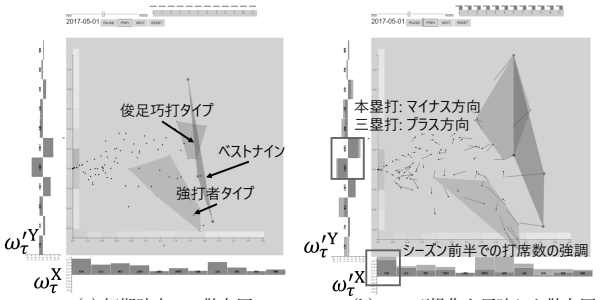
5 分析事例

分析対象には, プロ野球ヌルデータ[2]にて公開されていた, 日本プロ野球のパシフィック・リーグに所属する野手の 2017 年の打撃成績を用いる. シーズン全体で 80 打席以上出場した野手を選択し ($N = 98$), 本塁打や盗塁数などの成績を多次元属性 ($M = 11$) として, 15 日間隔でサンプリング ($T = 12$) したデータを分析に用いる.

可視化結果を図 4(a) に示す. 分析者は, 野手のシーズンを通した活躍度合いと打撃タイプの分類に興味を持ち, それらを評価する指標を作成したいと考えた. 最初に, 再生モードにて散布図の傾向を調べ, 時点によらず中央付近に出場回数が少ない選手, 右側には多い選手が

[1] http://flask.pocoo.org/

[2] 2017/11/8 公開終了



(a) 初期時点での散布図　(b) ユーザ操作を反映した散布図

図 4 野球データにおける分析事例

分布することを確認した. また, 右上には俊足巧打タイプ (打率が高く, 盗塁が多い), 右下には強打者タイプ (本塁打などの長打が多い) が多い傾向を確認した. これらの洞察に基づき, 分析者は X 軸を野手の活躍度合い, Y 軸を打撃タイプを表す指標として活用することにした.

次に, 分析者は静止モードで詳細な探索を行いながら, $\boldsymbol{\omega}$ を修正した. 例として, ベストナインを受賞するなど, シーズンを通して活躍していると判断した野手のグループについて作成した凸包に絶対的操作を行い, X 軸の大きい方にそれらの野手が配置されるように $\boldsymbol{\omega}$ を修正した. また, 各打撃タイプに該当すると判断した野手のグループごとに凸包を作成し, 相対的操作を用いてそれらが Y 軸の上下端に配置されるように $\boldsymbol{\alpha}$ を変更した後, 散布図ビュー内の UI を用いて明示的に $\boldsymbol{\alpha}$ の変更を $\boldsymbol{\omega}$ に反映した.

更新後の $\boldsymbol{\omega}$ (図 4(b)) を確認すると, X 軸における打席数はシーズン序盤の方が高く重み付けされているため, シーズン全体で安定した活躍をするためには序盤から試合に出る必要があると解釈した. Y 軸では, 単打や三塁打, 盗塁数が更新前よりもプラス方向に高く重み付けされているため. 俊足巧打タイプの野手の特性と解釈した. 二塁打や本塁打, 被四球は, 更新前後共にマイナス方向に重み付けされているため, 強打者タイプの特性と解釈した. また, $\boldsymbol{\omega}^{\mathbf{Y}}$ の時間的変化に関して, 四球はシーズン前半のほうが高く重み付けされる傾向が確認できた.

指標の叩き台となるパラメータ $\boldsymbol{\omega}$ の特性とドメイン知識の対応を確認した上で, 最終的に構築する指標を検討した. 例えば, 従来の塁打数[3]や出塁率[4]の計算式を参考に, 野手の活躍度合いを絶対値で, 特性を符号で表現する式 (3) のような指標 B_{type} を作成した. 分子に採用した属性と重みは Y 軸から得られた知見に基づいている. また, X 軸から得られた知見に基づき, シーズン前半では規定打席数を分母として打席数の傑出度合いを評価し, 後半では通常の出塁率と同様に, 各野手の打席数を分母とする.

$$B_{\mathrm{type}} = \begin{cases} \dfrac{\text{単打}\times 2+\text{盗塁数}\times 2+\text{三塁打}\times 2-(\text{四球}\times 2+\text{二塁打}+\text{本塁打}\times 3)}{\text{規定打席}} \\ (\text{シーズン前半}) \\ \dfrac{\text{単打}\times 2+\text{盗塁数}\times 2+\text{三塁打}\times 3-(\text{四球}+\text{二塁打}\times 2+\text{本塁打}\times 4)}{\text{打席}} \\ (\text{シーズン後半}) \end{cases} \tag{3}$$

6 おわりに

本稿では, 時系列データに対する評価指標の形成を支援するための視覚的分析フレームワークを提案した. そして, 実装したインタフェースを用いた実データの分析事例により, 提案フレームワークおよびシステムの有効性を示した. 今後は, ドメイン専門家によるケーススタディを通して, フレームワークの定性的な有効性検証を行う必要がある. また, 他の次元削減手法にも対応できるように, 提案フレームワークの拡張も検討する.

[3]塁打＝単打 × 1 + 二塁打 × 2 + 三塁打 × 3 + 本塁打 × 4
[4]出塁率 = (安打 + 四球 + 死球) ÷ (打数 + 四球 + 死球 + 犠飛)

参考文献

[1] D. A. Keim, F. Mansmann, J. Schneidewind et al.: Visual analytics: Scope and challenges, S. J. Simoff, M. H. Böhlen, A. Mazeika (eds.), Visual Data Mining, Springer, pp. 76-90, 2008.

[2] D. Sacha, M. Sedlmair, L. Zhang et al.: What you see is what you can change: Human-centered machine learning by interactive visualization, Neurocomputing, 2017.

[3] E. Horvitz: Principles of mixed-initiative user interfaces, Proc. of ACM SIGCHI Conference, pp.159-166, 1999.

[4] A. Endert, P. Fiaux and C. North: Semantic interaction for visual text analytics, Proc. of ACM SIGCHI Conference, pp. 473-482, 2012.

[5] R. Takami and Y. Takama: Visual analytics interface for time series data based on trajectory manipulation, Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence 2018 (WI'18), pp.342-347, 2018.

[6] H. Kim, J. Choo, H. Park, and A. Endert: InterAxis: Steering scatterplot axes via observation-level interaction, IEEE Trans. on Visualization and Computer Graphics, vol.22, no.1, pp.131-140, 2016.

[7] E. Wall, S. Das, R. Chawla et al.: Podium: Ranking data using mixed-initiative visual analytics, IEEE Trans. on visualization and computer graphics, Vol. 24, No. 1, pp. 288-297, 2018.

[8] W. Aigner, S. Miksch,W. Müller et al.: Visualizing time-oriented data - a systematic view, Computers & Graphics, vol.31, no.3, pp.401-409, 2007.

[9] A. C. Chen and X. Fu: Data + intuition: A hybrid approach to developing product north star metrics, Companion Proc. of WWW Conference pp. 617-625, 2017.

[10] J. Bernard, D. Sessler, T. Ruppert, et al.: User-based visual-interactive similarity definition for mixed data objects - concept and first implementation, Proc. of WSCG Conference, pp.329-338, 2014.