多次元データにおける複合インサイト探索の自動化

野澤　拓磨† 　董　　于洋† 　草野　元紀† 　小山田昌史†

† 日本電気株式会社 データサイエンス研究所　〒 211–8666 神奈川県川崎市中原区下沼部 1753

E-mail: †{nozawa-takuma,dongyuyang,g-kusano,oyamada}@nec.com

あらまし　本論文では, 多次元データにおけるインサイト探索を自動化し, 有用な可視化パターンを機械的に発見する手法を提案する．ルールベースのインサイト自動探索では, ユーザーが有用と考えるインサイトタイプを定義し, それらの顕著さを定量的に評価するスコア関数を用いてインサイトを与えるデータを選別する．複数のデータから構成されるインサイトは複合インサイトと呼ばれ, データ間の相関係数等に基づいてインサイトを評価する手法が先行研究で提案されている．しかしながら, 先行研究を用いて評価可能な複合インサイトは 2 次, すなわちペアワイズな関係性に限定されており, 高次の関係性を捉えることはできない．そこで本論文では, 上述の複合インサイトを拡張し, 任意の数のデータから構成される複合インサイトの評価手法を開発した．また, ダウ平均株価を用いた実験を行い, 提案手法によって類似した挙動を示すグループを抽出可能であることを示した．

キーワード　インサイト探索, 多次元データ, 可視化

1　はじめに

1.1　自動インサイト探索

近年, 様々な分野において蓄積されたデータを活用し, 人にとって有益な知見を見出す作業が行われている．このようにして発見された有益な知見はインサイトと呼ばれ, データ利活用シーンにおいてインサイトの発見は重要なタスクの 1 つであると位置づけられている．一般的なデータ分析作業においては, 分析者が仮説を設定し, 様々な角度から分析・可視化されたデータを元に, 仮説の検証を行うサイクルを繰り返す．当然ながら, 活用できるデータの増加に伴って様々なインサイトの発見が期待できるため, 多くのビジネスシーンでデータの蓄積や利活用を推進する動きがある．一方で, 分析対象の増大に伴い作業負荷も大きく増加するため, データ分析作業は非常に時間と労力を要するものとなっている．このような背景から, データからインサイトを自動的に発見するための研究に注目が集まっており,BI (Business Intelligence) ツールを中心に導入が進んでいる [1–3].

1.2　複合インサイト

多次元データセットにおける可視化パターンは, 下記の insight subject ごとに整理することが可能である．

$$insight\ subject = \{subspace, breakdown, measure\} \quad (1)$$

各変数の定義についての説明は省くが, 図 1 に示すように,subspace が凡例,breakdown が横軸,measure が縦軸に相当するものであり, これらの組み合わせごとに 1 つのデータが得られ, それをチャートとして可視化することができる．厳密に言えば,breakdown ごとの measure の集計方法（例：総和, 平均, 最大値, 最小値）についても指定する必要があるが, 簡単のため本論文では集計方法の記述を省略しており, 特に断りがない限り

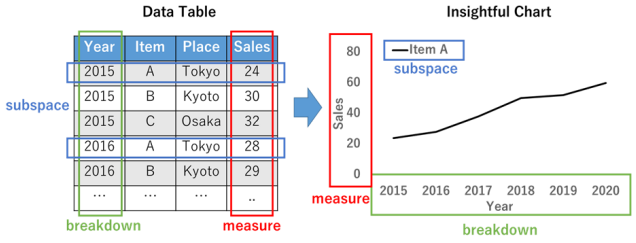

図 1　多次元データセットにおける可視化パターンの具体例

総和による集計を想定している．

上記のデータから得られるインサイトは, 入力変数となる insight subject の数に応じて Single Insight (シングルインサイト) と Compound Insight (複合インサイト) に分けることができる．シングルインサイトは入力変数である insight subject が 1 つの場合, すなわち 1 つのデータについて定義されるインサイトである．例えば, 図 1 の右図は, {Item A, Year, Sales} のトレンドに関するインサイトを示しており, シングルインサイトに分類される．一方で, 複合インサイトは複数の insight subject を入力変数にとるインサイトであり, 複数のデータの関係性を評価する．例えば,{Item A, Year, Sales} と {Item B, Year, Sales} の相関を評価するインサイトは複合インサイトに分類される．

複合インサイトの探索方法については,Tang et al. [2] や Ding et al. [3] によって研究がなされており, これまでに Correlation, Cross-measure correlation, 2D clustering の 3 つのインサイトタイプが提案されている．Correlation は,measure と breakdown が共通で,subspace が異なる insight subject によって与えられるインサイトであり, 似たような売上推移をした商品ペア等を発見することができる．Cross-measure correlation や 2D clustering は, subspace と breakdown が共通 で,measure が異なる場合のインサイトである．売上推移と相関が強いデー

タを Cross-measure correlation で発見し, その関係を示す散布図における外れ値を 2D clustering で発見することができる. データ分析において, 関係のあるデータを発見する作業は一般的に行われていることから, これらの探索を自動化することで, データ分析作業の効率化や新たなインサイトの発見に繋がることが期待できる.

1.3 本研究の目的

上記の先行研究で定義されている複合インサイトは, いずれも多数の可視化パターンの中から関係のある insight subject を発見するのに有用である. しかしながら, これらが捉えることができるインサイトは 2 次, すなわちペアワイズな関係性に限定されており, もっと複雑で高次な関係性が内在している可能性のあるビジネスデータへの適用には課題が存在する. 例えば, 商品の売上推移等は景気に大きく左右されるため, 商品データ全体の間に強い相関が見られることがある. この場合に, 上述の複合インサイトで各ペアの相関関係に基づいたインサイト探索を行ったとしても, 似たような傾向を示すデータやチャートが並ぶ結果となり, 背後にある全体的な関係を把握することは難しい. また, マーケター的な視点では, 景気が悪く全体的に売上が落ち込んでいる中で, 伸びている商品群が知りたいケースが考えられるが, 先行研究で提案されている手法では, このようなインサイトを発見することはできない. そこで本研究では,n 個 の insight subject の関係性, すなわち n 次の複合インサイトを主成分分析を用いて評価する手法を提案する. また, 類似したデータをクラスタリングによって纏めて取り扱うことで, 全体的な関係を俯瞰したインサイトを発見可能であることを示す.

本研究の貢献は以下の点にまとめられる.

（1） 先行研究において評価可能な複合インサイトが 2 次までに限定されていることを見出した.

（2） 主成分分析を用いて n 個の insight subject から得られる複合インサイト, すなわち n 次の複合インサイトを定量的に評価する方法を提案した.

（3） 疑似データ及びダウ平均株価の調整後終値データを使った実験を行い, 提案手法を用いて相関の強いグループを発見できることを示した.

（4） 上記の実験において, クラスタリングを用いて相関の強いデータを纏めて取り扱うことで, 全体の関係性を俯瞰したインサイトを捉えることが可能であると示した.

2 関 連 研 究

2.1 自動インサイト探索

自動インサイト探索の先行研究は大きく分けて knowledge-based 型,data-driven 型, 及びそれらのハイブリッド型の 3 つに大別される [1]. Knowledge-based 型は, 経験知に基づき人が有用と考えるインサイトの種類やその評価関数を事前に定義しておくことで, 目的のインサイトを与えるデータやチャートを選別する手法である [2,3]. 古くから数多くの研究がなされており, これまでに大きく分けて 12 タイプのインサイトに関する抽出法が提案されている [4]. Data-driven 型は, データテーブルとその可視化方法に関する情報を学習しておくことによって, 機械学習を用いた推論に基づき適切な可視化方法を提示する手法である [5]. 学習時には可視化方法についてのコーパスが必要になるため, データセットの整備に関する研究も並行して進められている [6]. ハイブリッド型は,data-driven 型の機械学習ベースの推論に加えて,knowledge-based 型の選別を取り入れた手法であり, 両者の利点を活かしつつ欠点を補うアプローチとして期待されている [7]. なお, 自動インサイト探索においては,knowledge-based 型のアプローチが比較的古くから研究されており, かつ現時点ではスタンダードな位置づけである.

2.2 インサイトスコア

Knowledge-based 型のインサイト探索において用いられる評価関数には様々な種類が存在する. 例えば,SeeDB [8], Zenvistage [9],VisPilot [10] では,Kullback–Leibler divergence や Earth Mover's distance 等を用いて 2 つのデータの分布の差を評価し, その差が大きい時にインサイトが得られるとみなしている. Foresight [11] や Voder [12] では, インサイトタイプごとに異なる基準を設けてその顕著さの評価を行っている. 複数のデータを分析する場合には,Scagnostics によって散布図の特徴を定量的に評価し, その特徴に応じてデータを分類する方法も提案されている [13,14]. また, 仮説検定の枠組みを用いた評価を導入している手法も存在し, この場合には異なるインサイトタイプのスコアを統一的なスケールで比較することが可能である [2,3,15,16].

2.3 仮説検定に基づくスコアの評価

前述のように,knowledge-based 型のインサイト探索においては, 人間が経験的に有用と考えるインサイト T を定量的に評価するスコア関数を導入することで, 多数の可視化パターンの中から有用なデータ, もしくは有用なデータを与える subspace, breakdown, measure の組み合わせを自動的に発見している. スコア関数の定義自体は各研究やインサイトタイプによって異なるが, 異なるインサイトタイプ間のスコアを比較したい場合には, 仮説検定の枠組みに基づいたスコア関数が導入される.

例えば, トレンドに関するインサイトは, 時系列データ $\boldsymbol{t} = \{(a_1, b_1), ..., (a_p, b_p) \mid a, b \in \mathbb{R}\}$ を直線回帰した際の傾き s^* がゼロであるとする帰無仮説を設定し,p 値を式 (2) のように計算する.

$$p_{\text{trend}} = \Pr(s > |s^*| \mid s \sim L(\mu, \delta)) \tag{2}$$

ここで,$L(\mu, \delta)$ はロジスティック分布であり, 先行研究 [2] においては $\mu = 0.2$ および $\delta = 2$ が用いられている. このとき, トレンドに関するインサイトのスコアは直線回帰の決定係数 r^2 を用いて 式 (3) のように与えられる.

$$score_{\text{trend}} = r^2(1 - p_{trend}) \tag{3}$$

同様に, 相関に関するインサイトの場合には,2 つの系列データ

$\boldsymbol{a} = \{a_1, ..., a_p \mid a \in \mathbb{R}\}$ と $\boldsymbol{b} = \{b_1, ..., b_p \mid b \in \mathbb{R}\}$ の Pearson 相関係数 ρ^* がゼロであるとする帰無仮説を設定し,p 値を式 (4) のように計算する.

$$p_{\text{correlation}} = \Pr(\rho > |\rho^*| \mid r \sim N(\mu, \delta)) \tag{4}$$

ここで,$N(\mu, \delta)$ は正規分布であり, 先行研究 [2] においては $\mu = 0$, $\delta = 0.05$ が用いられている. このとき, 相関に関するインサイトのスコア は式 (5) のように与えられる.

$$score_{\text{correlation}} = 1 - p_{\text{correlation}} \tag{5}$$

上記の式 (2) および式 (4) で与えられる p 値の定義域は $[0, 1]$ であり, 確率変数が顕著な値を取る場合, すなわちインサイトが得られるデータの場合には,p 値は 0 に近づくように小さくなる. このとき式 (3) および式 (5) から与えられるスコアは 1 に近づくように大きくなるため, インサイトタイプごとに用いられる確率変数が異なっていた場合でも, 統一的なスケールでスコアを比較することが可能である.

3 問 題 設 定

3.1 n 次の複合インサイト

Insight subject I が n 個与えられた場合の複合インサイトについて考える. それぞれの insight subject における subspace, breakdown, measure は本来は任意に選ぶことが可能であるが, 本研究では breakdown は n 個の insight subject で共通であるケースを想定する. 各 insight subject をそれぞれ添字で区別するように I_i と表記し, 一般化された n 次の複合インサイトのスコア関数 f_T を式 (6) のように記述する.

$$score = f_T(I_1, ..., I_n) \tag{6}$$

4 提 案 手 法

先行研究における複合インサイトは,$n = 1$ または 2 の場合に限定されているため, 式 (6) を既存技術で解くことはできない. また, 相関に関するインサイトの計算で用いられる Pearson 相関係数は 2 変数の場合にのみ定義される係数であり, これをそのまま n 変数に拡張することもできない. そこで, 本研究では相関行列を用いた主成分分析を行い,n 個のデータの相関に関する複合インサイトを評価する手法を提案する.

4.1 多次元データにおける相関の評価

主成分分析は, 多次元データにおける分散が最大となる軸をそのデータの特徴を表す主成分として求める手法であり, 第 1 主成分から順番に元のデータの情報をなるべく多く含むように合成されるので, 多次元データの次元縮約に用いられる.

相関行列を用いた主成分分析の概要を下記に示す. p 次元データ $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ を n 個を含む多次元データ $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)$ の相関行列は, 式 (7) のように表される.

$$\boldsymbol{R} = \begin{bmatrix} \rho(\boldsymbol{x}_1, \boldsymbol{x}_1) & \rho(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & \rho(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \rho(\boldsymbol{x}_2, \boldsymbol{x}_1) & \rho(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & \rho(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(\boldsymbol{x}_n, \boldsymbol{x}_1) & \rho(\boldsymbol{x}_n, \boldsymbol{x}_2) & \cdots & \rho(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \tag{7}$$

ここで,$\rho(\boldsymbol{x}_i, \boldsymbol{x}_j)$ は,$\boldsymbol{x}_i$ 及び $\boldsymbol{x}_j$ の Pearson 相関係数である. 最終的には, 主成分分析は式 (8) を満たす固有値 λ と固有ベクトル $\boldsymbol{u}$ を求める問題へと帰着される.

$$\boldsymbol{R}\boldsymbol{u} = \lambda \boldsymbol{u} \tag{8}$$

相関行列は半正定値性を満たすため,$n \times n$ の相関行列 $\boldsymbol{R}$ には n 個の非負の固有値が存在する. これらを大きい順番に $(\lambda_1 \geqq \lambda_2 \geqq \ldots \geqq \lambda_n)$ と並べたときの λ_i に対応する固有ベクトル $\boldsymbol{u}_i$ が第 i 主成分となる.

主成分分析で得られた固有値は各主成分における分散と対応する. 主成分分析では分散を情報として捉えているので, 各主成分の寄与率 $\lambda_i \ / \ \sum_{k=1}^{n} \lambda_k$ は各主成分の情報量を示す指標として用いることができる. すなわち, 相関の強い多次元データが与えられた場合には, それらを特徴づける軸が存在し, 各主成分の情報量すなわち寄与率に偏りが生じるため,n 個の insight subject の関係性を定量的に評価することができると考えられる. 得られた寄与率の分布やデータの性質からインサイトを定量的に評価する方法は複数考えられるが, 本研究では第 1 主成分の寄与率 $c_1^* = \lambda_1 \ / \ \sum_{k=1}^{n} \lambda_k$ を用いて, 式 (9) のように n 次の複合インサイトのスコアを計算する.

$$score = 1 - \Pr(c_1 > c_1^* \mid c_1 \sim L(\mu, \delta)) \tag{9}$$

ここで,$L(\mu, \delta)$ はロジスティック分布であり, 本研究では $\mu = 0.5$ および $\delta = 0.1$ を採用している. 帰無仮説で用いる分布やパラメータの選定には任意性があるが, ここでは第 1 主成分のもつ情報量が全体のうち 50 % 以上を占めるような場合に, その程度に応じてスコアが高くなるように設定している.

4.2 n 次の複合インサイトの計算コスト

Tang et al. (2017) [2] で言及されているように, シングルインサイトの計算コストは探索する insight subject の数に比例する. Subspace の数を N_s,breakdown として選択可能な列数を N_d,measure として選択可能な列数を N_m, シングルインサイト におけるインサイトタイプの数を $|T_1|$ とすると, シングルインサイト における計算コストは, 式 (10) のように表される.

$$O(N_s \cdot N_d \cdot N_m \cdot |T_1|) \tag{10}$$

次に,n 次の 複合インサイト, すなわち insight subject が n 個与えられる場合を考える. このとき subspace 及び measure の選び方はそれぞれ ${}_{N_s}C_n$, ${}_{N_m}C_n$ 通り存在するため,n 次のインサイトタイプの数を $|T_n|$ とすると, その計算コストは式 (11) のように表される.

$$O\left(\left(\frac{N_s!}{n!(N_s - n)!}\right) \cdot N_d \cdot \left(\frac{N_m!}{n!(N_m - n)!}\right) \cdot |T_n|\right) \tag{11}$$

もし,n が定まっていない場合には,2 次から n 次までのインサイトを総当りで計算する必要がある．二項係数の総和は $\sum_{k=0}^{n} {}_nC_k = 2^n$ であることを利用すると, その計算コストは式 (12) のように表される．

$$O((2^{N_s} - N_s - 1) \cdot N_d \cdot (2^{N_m} - N_m) \cdot |T_n|) \qquad (12)$$

4.3 クラスタリングによる計算効率化

式 (12) が示すように,n 次の複合インサイトの計算コストは N_s 及び N_m の増加に伴い指数関数的に増大してしまうため, 大規模なデータセットを扱うには都合が悪い．これは,n 個の insight subject の組み合わせ爆発に起因しており, 上述の主成分分析を用いたインサイト探索だけではなく,n 次の複合インサイトの計算全般に付随する問題である．例えば, 実験 5.2 で扱う問題設定のサイズ $(N_s = 10, N_d = 1, N_m = 1)$ の insight subject の組み合わせ数は 1,013 通りだが, 少しデータセットを増やして $(N_s = 30, N_d = 1, N_m = 1)$ とすると,1,073,741,793 通りの組み合わせを考慮する必要があり, 現実的な計算時間ではなくなってしまう．

そこで本研究では, データセット中に非常に強い相関を持つグループが存在する場合には, 相関の強いグループを 1 つに纏めたクラスタ単位で主成分分析を行う方法を提案する．これによって, 考慮しなければならない組み合わせの数を大幅に削減し, 計算コストを現実的なレベルまで低減することができる．また,5.2.1 の結果からも明らかなように, 相関の強いグループを構成しているデータのスコアは高くなる傾向があるため, 単純にスコアの Top-k を表示したとしても, どれも類似した傾向を示すチャートばかりとなってしまい, 情報量の割に有用な示唆が得られない場合がある．上述のクラスタ単位で粗視化した分析は, データの関係性を俯瞰したインサイトをユーザーに提示する効果も期待できるため,n 次の複合インサイトの分析に有用な手法であると考えられる．

データのクラスタリング方法やその纏め方については議論の余地がある．例えば, データをどのような単位や粒度で纏めるかについては, データの性質と分析の目的次第で調整が必要になると考えられる．また, データを纏める際の集計方法についても, 幾つかの選択肢が存在する．最終的にはこれらの部分についても自動的に最良と思われるクラスタリング手法や集計方法を探索できるようにすることが望ましいが, 本研究では実験的に Ward 法を用いた階層的クラスタリングを行い, 各クラスターの中で最もスケールが大きいデータを代表点とすることで, 相関の強いデータを纏めて取り扱っている．上記の詳細については,5.2.2 で述べる．

5 実　　験

5.1 擬似データ

まず最初に簡単に分析可能な擬似データを用いて, 提案手法の有効性を示す．図 2 の (a) は一様分布,(b) は幾何ブラウン運動によって生成したデータであり, それぞれ異なる 10 個の時系列を含む．(c) は (b) のうち相関が大きい 6 つの時系列を取り出したものである．一様分布によって生成された (a) の場合には, 各主成分の寄与率はほぼフラットな分布となっているが, これは 10 個の時系列がほぼ無相関であるために, どの向きに軸をとっても分散が一定であることに起因している．一方で, 幾何ブラウン運動によって生成された (b) の場合には, 第 1 主成分の寄与率が明らかに突出している．これは 10 個の時系列の中には相関が強いデータが存在するため, それらを特徴づける主成分, すなわち分散が大きくなる軸が存在することを示している．(b) のデータのうち相関が強いグループを取り出した (c) の場合には, この主成分はより顕著になるため第 1 主成分の寄与率はさらに大きくなる．これらの結果は, 第 1 主成分の寄与率を n 次の複合インサイトスコアの計算に用いることで, 有益なチャートを与える insight subject の組み合わせを発見可能であることを示している．

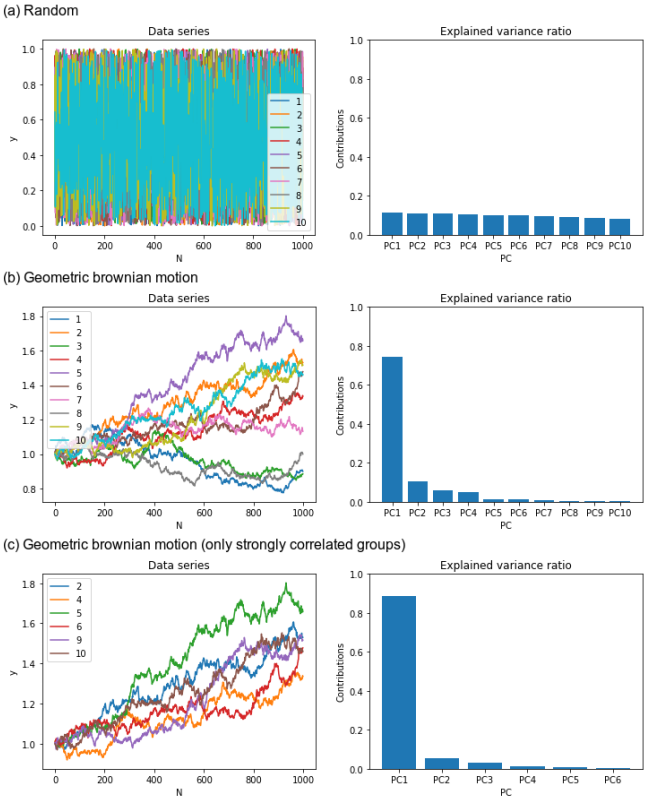

図 2　時系列データ（左図）と各主成分の寄与率（右図）

5.2 株 価 推 移

5.2.1 全探索を行った場合

次に, 実際のビジネスデータにおける有効性を検証するために, ダウ平均株価の構成銘柄の中から無作為に選んだ 10 銘柄の株価データの分析を行う．2019 年 1 月 1 日から 2019 年 12 月 31 日までの調整後終値のデータを入力としてそれらの全組み合わせについて分析を行い, 第 1 主成分の寄与率の大きい順番に Top 10 まで並べると表 1 のようになる．本実験では, 各 insight subject における breakdown を日付,measure を調整後終値としているので,subspace の組み合わせのみを考慮している．図 3 は, 表 1 における Top 1, 4, 7, 10 を与えるデータ

セットを示しているが, 同様の傾向を持ったグループを抽出できていることが分かる.

表 1　全探索した場合の Top 10 インサイト

rank	score	insight subjects (subspaces)
1	0.991911	'PG', 'WMT'
2	0.991384	'MSFT', 'PG'
3	0.990856	'MSFT', 'WMT'
4	0.990644	'MSFT', 'PG', 'WMT'
5	0.984714	'GS', 'MSFT'
6	0.984469	'GS', 'NKE'
7	0.983697	'GS', 'MSFT', 'PG', 'WMT'
8	0.983100	'GS', 'MSFT', 'PG'
9	0.982725	'GS', 'MSFT', 'WMT'
10	0.982231	'GS', 'PG', 'WMT'

図 3　表 1 において Top 1, 4, 7, 10 を与えるデータセット

5. 2. 2　クラスタ単位で分析を行った場合

5. 2. 1 では, 提案手法が相関の強いグループを自動的に発見することを確認できた. しかしながら, 表 1 を見ると Top 10 の insight subjects は非常に類似していることが分かる. これは, データセット中に互いに強い相関を持つグループが存在した場合には, そのグループを構成するデータの組み合わせも同様のスコアとなってしまうことに起因する. この場合, スコアが高くなる insight subject の組み合わせを順番にユーザーに提示したとしても, 類似したチャートばかりとなってしまい情報量の割に有益な示唆が得られにくい. また, データセットの数が増えるとその組み合わせが爆発的に増えてしまうことも問題である.

上記の課題を解決するため,4. 3 のように相関の強いグループを 1 つに纏めたクラスタ単位で主成分分析を行った場合の結果を示す. 図 4 は,1 から Pearson 相関係数を引いたものをデータ間の距離として Ward 法で樹形図を作成したものである. Ward 法では, この樹形図によってクラスタを分割する閾値を決定するが, ここでは特に強い相関が見られる ('MSFT', 'PG', 'WMT') と ('GS', 'NKE') を纏めて取り扱うことにする. その後, 各クラスタにおいて最もスケールが大きいデータを代表点とすることで, 相関の強いグループを纏めて取り扱う. なお,

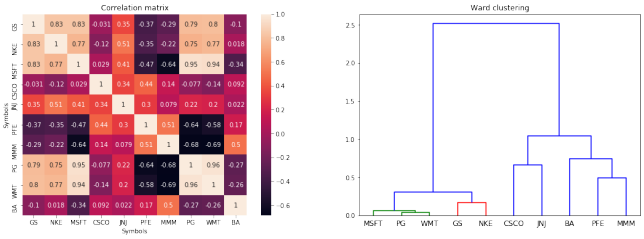

図 4　時系列データ間の Pearson 相関係数（左図）に基づき Ward 法でクラスタリングを行った結果（右図）

表 2　クラスタ単位で分析した Top 10 インサイト

rank	score	insight subjects (subspaces)
1	0.984714	('MSFT' 'PG' 'WMT'), ('GS' 'NKE')
2	0.961412	('MSFT' 'PG' 'WMT'), ('MMM')
3	0.926559	('PFE'), ('MMM')
4	0.923475	('MMM'), ('BA')
5	0.913275	('MSFT', 'PG', 'WMT'), ('PFE')
6	0.912112	('MSFT' 'PG' 'WMT'), ('GS' 'NKE'), ('MMM')
7	0.901019	('CSCO'), ('PFE')
8	0.895809	('MSFT', 'PG', 'WMT'), ('GS', 'NKE'), ('PFE')
9	0.885627	('MSFT', 'PG', 'WMT'), ('JNJ')
10	0.880063	('MSFT', 'PG', 'WMT'), ('GS', 'NKE'), ('JNJ')

階層的クラスタリングを用いた理由は, 結果の再現性が担保されており, データ全体の関係性を俯瞰しながらクラスタリングの粒度が選択可能という点が実験的な導入に適しているためである.

クラスタリング後に主成分分析を行い, 第 1 主成分の寄与率の大きい順番に Top 10 まで並べると, 表 2 のようになる. スコアが最大となるのは, 表 1 の Top 10 を構成している ('MSFT', 'PG', 'WMT', 'GS', 'NKE') が subspace に与えられたときであり, 上記の手法を用いることで個々の時系列データの関係性を俯瞰したインサイトを捉えることができている. 図 5 は, 表 2 における Top 1 を与えるデータセットを示しているが, クラスタリングを行わない場合と同様のインサイトを抽出できていることが分かる. なお, 元の時系列データにおいては, ('MSFT', 'PG', 'WMT') と ('GS', 'NKE') のグループの間に特徴的な差がないように見えるが, 標準化したスケールで比較するとそれぞれのグループの特徴を捉えた分け方になっていることが分かる. 図 6 は, 表 2 における Top 2 を与えるデータセットを示している. ('MMM') は ('MSFT', 'PG', 'WMT') のグループとは部分的に異なる動きをしつつも, 全体的には似た傾向を示していることが分かる. 表 1 のリストには ('MMM') を含むグループが現れていないが, これは ('MSFT', 'PG', 'WMT') や ('GS', 'NKE') を構成するデータの組み合わせが表 1 のランキング上位を占めてしまっていることに起因している. すなわち, クラスタリングを行うことで傾向の近いデータの情報が纏められ, 表 2 においては全体を俯瞰したインサイトが得られるようになっていると言える.

計算量の削減の観点からもクラスタリングの恩恵は大きい. 表 5. 2. 2 は, クラスタリングの有無による,insight subject の組み合わせ数, すなわち n 次の複合インサイトの計算コストの変

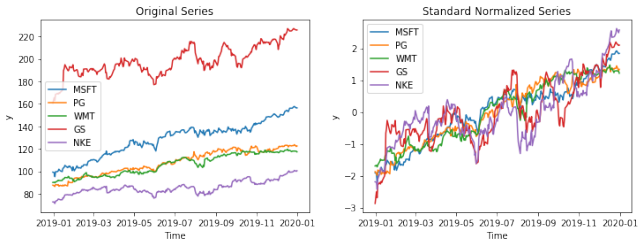

図 5　表 2 において Top 1 を与えるデータセット

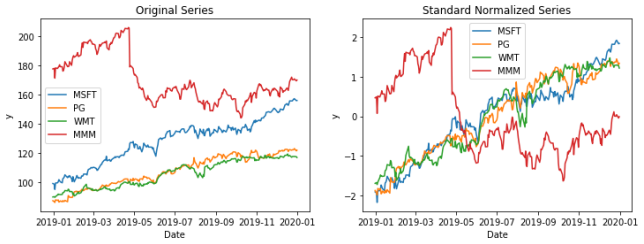

図 6　表 2 において Top 2 を与えるデータセット

表 3　クラスタリングの有無による計算コストの変化

	10 銘柄	30 銘柄
クラスタリングなし	1,013	1,073,741,793
クラスタリングあり	120	262,125

化について比較したものである．ここでは, ダウ平均株価を無作為に 10 銘柄選んだ場合に加えて,30 銘柄全てを使用した場合についても示しているが, データセット以外の設定は同じものを用いている．10 銘柄のデータセットにおける計算コストは, クラスタリングによって約 1/8 に,30 銘柄のデータセットの場合には約 1/4000 へと削減できている．このように, データセットの大規模化に伴いクラスタリングによる計算量の削減効果も大きくなるため, 大規模データセットにおける組み合わせ爆発をある程度緩和することが期待できる．

クラスタリングの粒度については, データの性質と分析の目的次第で調整が必要となる部分であるが, 過度に纏め過ぎずある程度特徴的な性質を持った単位で分けられていれば, 表 2 の Top 1 のように 似た傾向を持ったグループ同士は高いスコアで評価されるため, 結果的に n 次の複合インサイトの insight subject という形で特徴的なグループが抽出される．また, クラスタリングを行うことで, 表 1 のように個々のデータ単位でのインサイトが失われてしまう問題については, まず特徴的なインサイトを示すクラスタ単位のインサイトを提示した後, 必要に応じてそれらを構成する個々のデータレベルのインサイトを計算することで解決可能である．

5.3　ま　と　め

5.1 および 5.2 では, 提案手法の有効性を示すため疑似データとダウ平均株価を使った実験を行った．図 2 が示すように,n 個のデータの相関の強さは主成分分析における第 1 主成分分析の寄与率の大きさに反映される．これを式 (9) のように, 帰無仮説の枠組みに従ってスコア化したものを用いると, 表 1 のように相関の強いデータのグループを自動的に発見することが可能である．あるグループを構成するデータの組み合わせによって, スコアのランキング上位が埋まってしまう問題については, 予め相関が強いグループをクラスタリングして纏めておくことで, 表 2 のように全体を俯瞰したインサイトを得ることが可能である．

6　結　　論

本研究では, 主成分分析における各主成分の寄与率をインサイトを表す指標として用いることで, データセットにおける n 次のインサイトを定量的に評価する手法を提案した．これによって, これまでユーザーが認識することが難しかったデータ間の高次な関係を自動的に発見することが可能になった．また,n 次のインサイトには, それを構成する insight subject の組み合わせの数が爆発的に増えてしまう問題があるが, 本研究では相関の強いデータを纏めたクラスタ単位で分析を行うことで, 計算コストの飛躍的な増大を軽減し, さらにデータの関係性を俯瞰して捉えることを可能にした．なお, 有効性の検証のために行った実験 5.2.1 及び 5.2.2 はどちらも時系列データに関するものであるが, 提案手法は n 個の insight subject で共通した breakdown が存在しデータ間の相関行列が計算できる場合であれば, 横軸がカテゴリ変数のような集計データにもそのまま適用可能である．

文　　献

[1] Sujia Zhu, Guodao Sun, Qi Jiang, Meng Zha, and Ronghua Liang. A survey on automatic infographics and visualization recommendations. *Visual Informatics*, Vol. 4, No. 3, pp. 24–40, 2020.

[2] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. Extracting top-k insights from multi-dimensional data. In *Proceedings of the ACM International Conference on Management of Data*, pp. 1509–1524, 2017.

[3] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. Quickinsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of the ACM International Conference on Management of Data*, pp. 317–332, 2019.

[4] Po-Ming Law, Alex Endert, and John Stasko. Characterizing automated data insights. In *Proceedings of the IEEE Visualization Conference*, 2020.

[5] Kevin Hu, Michiel A Bakker, Stephen Li, Tim Kraska, and César Hidalgo. Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[6] Kevin Hu, Snehalkumar'Neil'S Gaikwad, Madelon Hulsebos, Michiel A Bakker, Emanuel Zgraggen, César Hidalgo, Tim Kraska, Guoliang Li, Arvind Satyanarayan, and Çağatay Demiralp. Viznet: Towards a large-scale visualization learning and benchmarking repository. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[7] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. Deepeye: Towards automatic data visualization. In *Proceedings of IEEE International Conference on Data Engineering*, pp. 101–112. IEEE, 2018.

[8] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. Seedb: Efficient

data-driven visualization recommendations to support visual analytics. In *Proceedings of the VLDB Endowment*, Vol. 8, p. 2182, 2015.

[9] Tarique Siddiqui, Albert Kim, John Lee, Karrie Karahalios, and Aditya Parameswaran. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. In *Proceedings of the VLDB Endowment*, Vol. 10, pp. 457–468, 2016.

[10] Doris Jung-Lin Lee, Himel Dev, Huizi Hu, Hazem Elmeleegy, and Aditya Parameswaran. Avoiding drill-down fallacies with vispilot: assisted exploration of data subsets. In *Proceedings of the International Conference on Intelligent User Interfaces*, pp. 186–196, 2019.

[11] Çağatay Demiralp, Peter J Haas, Srinivasan Parthasarathy, and Tejaswini Pedapati. Foresight: Recommending visual insights. In *Proceedings of the VLDB Endowment*, Vol. 10, pp. 1937–1940, 2017.

[12] Arjun Srinivasan, Steven M Drucker, Alex Endert, and John Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 25, No. 1, pp. 672–681, 2018.

[13] Graham Wills and Leland Wilkinson. Autovis: automatic visualization. *Information Visualization*, Vol. 9, No. 1, pp. 47–69, 2010.

[14] Tuan Nhon Dang and Leland Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *Proceedings of the IEEE Pacific visualization symposium*, pp. 73–80. IEEE, 2014.

[15] Cuiyun Gao, Jichuan Zeng, David Lo, Chin-Yew Lin, Michael R Lyu, and Irwin King. Infar: Insight extraction from app reviews. In *Proceedings of the ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 904–907, 2018.

[16] Yohan Bae, Suyeong Lee, and Yeonghun Nam. Timesight: Discovering time-driven insights automatically and fairly. In *Proceedings of the International Conference on Software Engineering and Knowledge Engineering*, pp. 49–54, 2020.