情報処理学会論文誌　データベース　Vol. 1　No. 2　15–25 (Sep. 2008)

業務データ分析のための データ分析フレームワークの開発

末永高志[†1]　山中啓之[†2]　高橋彰子[†1]
東陽子[†1]　佐治美歩[†1]　矢野順子[†1]
中川慶一郎[†1]　関根純[†1]

業務プロセスの改善，経営課題の解決，事業の業績管理といった企業活動の合理化を促進するための，データ分析技術が注目されている．本稿では，企業におけるデータ分析の導入に必要なデータ分析の試行評価業務において，データ分析手順に関するノウハウを蓄積するための，データ分析フレームワークを提案する．これは，約 100 件のデータ分析経験をもとに，データ分析のノウハウを目的別に分類した 9 つの分析シナリオで構成されたものである．既存のアプローチでは，業務分野に特化したソリューションを用意するという方法がとられていたが，我々のフレームワークでは，目的指向で分類しているため業務目的が同じであれば業務分野が異なっても適用可能である．これにより，データ分析のノウハウを共有することが可能となる．当フレームワークの効果を検証するためにデータ分析作業の実験を行い，データ分析作業の経験がある場合は，分析報告書の品質の約 67%向上に貢献することと，作業時間の約 64%を削減できることを確認した．

A Framework for Business Data Analyses

Takashi Suenaga,[†1] Hiroyuki Yamanaka,[†2]
Shoko Takahashi,[†1] Yoko Higashi,[†1] Miho Saji,[†1]
Junko Yano,[†1] Kei-ichiro Nakagawa[†1]
and Jun Sekine[†1]

Data analysis has become a key technology in achieving business process re-engineering, solving corporate management issues, or promoting services, based on quantitative monitoring of business activities. This paper proposes a framework for business data analysis. It consists of 9 typical data analysis scenarios classified according to business goals, which are extracted from about 100 real cases. It is different from conventional domain specific data analysis solutions in that it can be applicable to different business domains as far as business goal is the same. It is useful when data analysts have to share data analysis know-how. Experiments show that the quality of analysis increases by 67% and the time for data analysis decreases by 64% for experienced data analysts.

1. はじめに

企業のシステムに蓄積されたデータを分析し，経営課題の解決や，日々の業務の効率化を目指すビジネスインテリジェンスという概念[1] の重要性が広く認知されている．

この実現のために OLAP（on-line analytical processing）[2] やデータマイニング[3),4)] と呼ばれる概念が提唱され，様々な実用化がなされてきた．OLAP は多次元データ分析と呼ばれ，様々な属性（性別，年齢，地域など）や時間軸ごとのデータの要約，比較を行うことで，様々な気づきを与えるものである．また，データマイニングは，小売店の併売分析などによく使われ，POS データといった大量に蓄積されたデータの中から，よく見られるパターンを抽出する技術である．これらの技術で実現されることは，蓄積された業務データの「見える化[5]」であるが，データ分析を行うまで実際に役に立つ情報が得られるかが分からないため，分析システム導入の投資対効果が分かりにくい．このため，分析システムの導入に先立ち，現実に蓄積されたデータからどのような経営課題が解決できるかを，データ分析のトライアルを通じて実証することの重要性が増してきた．

我々はこれまでこのトライアル活動を実施してきたが，その際に，データ分析のノウハウが属人的にしか蓄積されないという課題があった．具体的には，異なる担当者が，異なる業務分野で，同じような分析を別々に行っているなど，非効率な状況が見られてきた．ここで，データ分析を行うにあたって必要となるノウハウとは，たとえば，統計的な手法で予測式などのモデル式を作成したのち，「予測式をもとにして投資対効果を最大化する」といった，データ分析の手順やその活用方法に関するノウハウである．

このような課題に対して，ノウハウを組織的に活用するために，主に 2 つの方法が提案されている．1 つは，データ分析の処理フローを可視化するツール[6)–9)] で，もう 1 つが，CRM などの業務分野を特定した分析のテンプレート[10] を用意しておく方法である．これらは，

†1 株式会社 NTT データ技術開発本部
Research and Development Headquarters, NTT Data Corporation
†2 株式会社 NTT データビジネスソリューション事業本部
Business Solutions Sector, NTT Data Corporation

いずれもデータ分析の手順を保存するものとなっている．

データ分析の処理フローを可視化するツールによって，過去に作成したデータ分析の手順を管理しやすくなったが，事例に特化した処理フローを残すだけでは再利用が困難である．データ分析を行う手順を大まかに分けると，データから把握できることを抽象化するモデル構築の処理と，そのモデルを業務に活用するための処理が必要である．しかしながら，データに応じて使用する適切な分析手法やデータの構成は，事例やデータの種類によって異なるため，実際のデータや利用した分析手法が処理フローとして残されているだけでは，どの処理を意図したものか分かりにくい．

一方で，業務分野を特定して分析手順をテンプレート化する方法では，分析の活用方法やデータの構成がある程度規定されるため，利用する分析ツールやデータベース構造などをテンプレートに組み込むことができ，トライアル完了後のシステム導入が短期間で可能となる．しかしながら，異なる業務分野に転用することは考えられておらず，限定的な範囲での対応しか期待できない．

以上から，データ分析のノウハウを組織に蓄積するためには，いかに業務分野を限定しない形式で，分析手順や活用方法のノウハウを形式知化するかがポイントである．この課題に対して我々は，様々な業務分野で行われたおよそ 100 件のデータ分析の事例をもとにデータ分析のノウハウを蓄積する方法を検討した結果，業務分野単位ではなく，業務目的単位でのノウハウ化に着目した．

そこで本稿では，データ分析を業務分野ではなく業務目的別に類型化し，それぞれに対して分析手順をシナリオとしてまとめることで，業務分野を限定せずデータ分析に必要な様々なノウハウを蓄積することのできる，データ分析フレームワークを提案する．

また，提案するフレームワークの効果を検証するために，データ分析業務の経験のない被験者に対していくつかの課題を出し，フレームワークに則って作成された分析シナリオを利用した場合の，作業時間の短縮度合いや分析報告書などの品質を，実験により定量的に評価する．

2. データ分析フレームワーク

本章では，データ分析フレームワークの構築に必要な分析目的の類型化について議論し，さらにそれに対応する分析シナリオの構成と記述上の工夫点を述べる．まず，2.1 節では過去のデータ分析の事例をもとに，データ分析を目的別に類型化する考え方を示す．次に，2.2 節では類型化された分析目的に対して，分析の手順をシナリオとしてまとめるための，

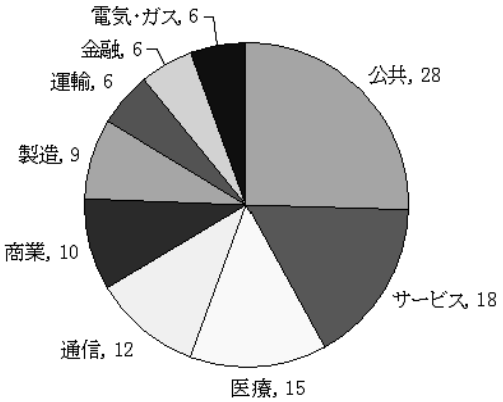


図 1　過去のデータ分析事例の業務分野別件数
Fig. 1　Distribution of data analysis cases over business domains.

記述すべき項目について整理する．また，2.3 節では提案するフレームワークに則って記述した，分析シナリオの具体例を示し，2.4 節では作成した分析シナリオの活用方法について述べる．

2.1　分析目的の類型化

データ分析のノウハウを蓄積するにあたって，過去に関わった様々な業務分野にまたがる 100 件ほどの実績（**図 1** を参照）を整理したところ，使われているデータ分析手法は様々なものが見られる一方で，異なる業務分野間であっても分析の課題とデータ構造が類似している事例があることに着目した．そこで本節では，2 つの具体的な事例をもとに，それらがデータ分析の目的という観点で類型化できることを示す．

まず，小売店が販売促進活動を行う例を考える．販売促進のためには顧客に商品を認知してもらう必要があるが，広告やダイレクトメールといった活動には必ずコストがかかる．このような問題に対して，効率的に売上げを伸ばすような販売促進活動を実現するためには，データ分析によって活動のコストとリターンの関係を定量化したうえで，(1) どのような販売促進の手段を実施し，(2) ターゲットとする顧客をどのような選出基準で何人選び出せばいいのか，を明確化しなくてはならない．データとしては，顧客ごとの属性情報や購入履歴データを活用することになる．これらのデータ分析から，商品購入につながる各種要因を明確化し，販売促進手段やターゲット顧客の選び方を決定する．さらに，ターゲット顧客の人数を変更して損益を試算することで，適切な施策を選択することが期待される．

表 1　分析シナリオと分析事例
Table 1　Data analysis scenarios and their examples.

分析シナリオ　　サブシナリオ	概要	分野数
(1) 評価・要因分析型	様々な対象の比較評価と改善要因の特定	7
(2) 異常値検出型　不正検出型	不正・異常の定義と合致/類似する行動状態の検出	1
外れ値検出型	標準的な行動・状態の定義と逸脱の検出	2
(3) コンテクスト・アウェアネス型	行動履歴・嗜好の分析から一歩先回りしたサービスの提示	1
(4) 予兆発見型	行動変化や状態変化の監視による予兆の発見	1
(5) 予測・制御型　収益シミュレーション型	業務効率化による増収効果の試算	6
リスク・シミュレーション型	業務のモデル化と不確実要素によるリスクの試算	2
リスク・ヘッジ型	業務のモデル化とリスク分散手法を用いたリスク低減策の提示	2
最適化型	業務のモデル化と最適化手法を用いた意志決定策の提示	6
(6) ターゲティング型	見込み顧客など重点アプローチすべきターゲットの抽出	3
(7) プロセス・トレース型	成長・発展プロセスの抽出と促進/阻害要因の特定	4
(8) 与信管理型	顧客・企業の滞納・倒産リスクの試算と融資判断の支援	2
(9) マーチャンダイジング型	様々な視点での売れ筋ランクの作成と品揃えの決定	2

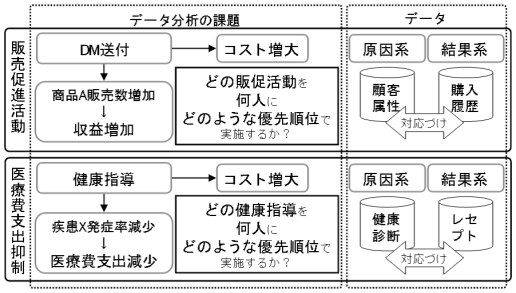


図 2　異なる業務分野間での類似したデータ分析の構造
Fig. 2　Similarity of data analysis flows across different business domains.

次に，医療費支出の抑制が課題となっている保険者（健康保健組合や保険会社など）の例を考える．保険者の中では，発症する前にカウンセリングや禁煙指導・食事指導などの実施により疾病の発症率を低め，結果的に医療費支出を抑制しようとする動きが活発化しつつある．健康指導によって発症率が低下し医療費支出が減ることが期待されるが，すべての被保険者に健康指導を実施するとコストもかかってしまう．そのため，(1) どのような健康指導を実施し，(2) ターゲットとする被保険者をどのような選出基準で何人選び出すのか，を決定しなくてはならない．この場合は，被保険者ごとの直近の健康診断情報や，レセプトデータ[⋆1]から得られる発症履歴を活用することになる．これらのデータ分析から，発症リスクや損益に関わる各種要因を明確化し，健康指導項目や健康指導の対象とする被保険者の選び方を決定する．ここから，選び出す被保険者の数を変更することで損益を試算し，適した施策を明確化する．

これら 2 つの分析例を対応づけたものを**図 2** に示す．この図から，データ分析の課題では「DM 送付」や「健康指導」というコストのかかるアクションや，「収益増加」や「医療費支出減少」という目的が対応しており，また取り扱うデータの構造では，原因系のデータが「顧客属性」と「健康診断」，結果系のデータが「購入履歴」と「レセプト」という対応がつくことが分かる．このように，小売業の販売促進活動と保険者の発症防止活動とは，課題・データの構造が非常に似ている．この分析の課題は，アクションに対するコストとリターンの関係から，収益を最大化するようなアクションをシミュレーションによって選択することから，「収益シミュレーション型」と命名した．これは，業務改善を行った結果の収益効果を試算することを目的としている．

このように，類似する事例を整理することを行い，**表 1** のとおり 9 つの分析シナリオにまとめた．さらに，分析シナリオによってはサブシナリオとしてさらに細分化できるものが

⋆1 患者が受けた診療に対して，医療機関が健保組合などの保険者に請求する医療費の明細書のことで，診療や処方した薬の費用が記載されているものである．

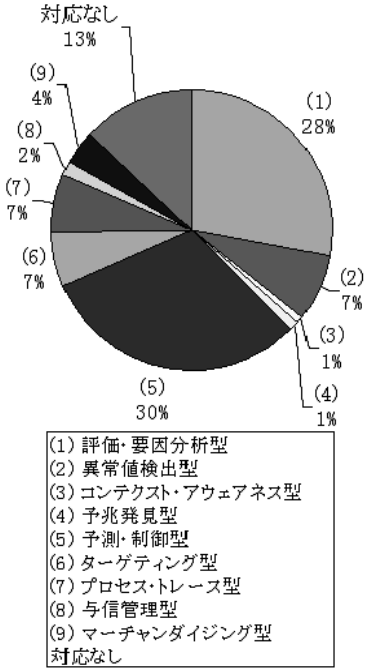


図 3　各々の分析シナリオで対応した分析事例の割合
Fig. 3　Number of cases for each scenario.

あり，それらを合わせると 13 の分析シナリオとなった．また，分野数とは各々の分析シナリオで対応した分析事例の業務分野の数を示したものである．ここで，(3) や (4) の分析シナリオは分野数が 1 となっているが，これらはもともと事例数が少なかったためである．また，(5) は事例数が多い分析シナリオであり，このような分析シナリオはサブシナリオに細分化されている．

さらに，**図 3** に分析シナリオで対応できた事例の割合を示す．今回の提案で全体の 87%の事例に対応できていることが分かる．ここで，対応できなかった事例とは OLAP ツールやデータマイニング・ツールの選定，質問票に関するテキスト・データ分析などがある．

2.2　分析シナリオの記述項目

類型化された分析目的には，それぞれに特徴的なデータ分析のための課題や手順などがある．これらのノウハウを蓄積し共有するためには，分析シナリオ間での統一的な記述が重要である．そこで，記述すべき項目を**表 2** のようにまとめた．

表 2　ノウハウの蓄積のために記述する項目
Table 2　Data items describing data analysis scenarios.

観点	項目
理解を助けるための情報	分析の目的 事例 訴求ポイント
分析の手順	モデルの構築 モデルの活用
分析手法の選択肢	適用できる分析手法

分析シナリオに最初に求められることは，分析者や顧客がシナリオの内容を理解することである．ここで，内容の理解に必要な情報は対象者によって異なり，様々な業務分野に対してデータ分析を行う分析者には，より抽象化された表現によって記述されていることが必要である．一方で，分析内容を顧客に説明する場合は，抽象化された表現よりもむしろ具体的な例で説明する方が理解を促す傾向がある．さらに，顧客に興味を持ってもらうためには，「ターゲット顧客の抽出」や，「コストとリターンの最適配分」といった，データ分析によって何を達成するものかを具体的に説明したほうがより訴求しやすい．したがって，分析者のための「分析の目的」の記述と，顧客の理解を促進するための，「事例」や「訴求ポイント」を記述することとした．

分析の内容を理解したら，次はデータ分析の手順を理解する必要がある．そこで，我々のフレームワークの着目点である目的別の類型化に則って，データ解析手法による統計モデルや最適化モデルを構築する「モデルの構築」の手順に加え，作成したモデルを活用するための「モデルの活用」の手順を記述することとした．

また，各々の手順では，いくつかの分析手法の中から適切なものを選択しなければならない．たとえば，収益シミュレーション型でモデルを構築する場合はよく回帰分析[11] と呼ばれる分析手法が使われるが，モデルの目的変数や説明変数に使うデータ項目の値が連続値のみで構成される場合は重回帰分析が適用でき，さらに，離散的な値が含まれる場合はロジスティック回帰分析が適切であったり，そのほか，購入の決定有無のルールを可視化したいといった要望がある場合は，精度に関してはやや劣るものの決定木分析[12] の適用を検討したりと，状況によって適切な手法が異なる．そこで，それぞれの手順に対して分析手法の選択肢をまとめるために，「適用できる分析手法」を記述することにした．

これらを具体化するための手順は以下のとおりである．まず，特定の分析シナリオに対応づけられた過去の事例をもとに，特に業務分野の個別性の高いデータ項目の選択処理や外れ値除去といった処理を，前処理部としてまとめる．そして，分析処理部に対応するデータや分析手法をまとめ抽象化する．具体的には，モデルの構築に用いる実際のデータの項目名や，事例によって異なるデータの項目数を隠蔽するために，説明変数や目的変数という項目名にまとめる．また，一般的に分析手法で構築したモデルは検証するための手順が欠かせないのだが，そのために用意したデータを検証用データと明記する．さらに，分析手法の処理フローに関しては，モデルを構築する手順とモデルを活用する手順にそれぞれまとめる．これらを行ったのち，各々の手順に対して階層的な構成にし，上位の階層では手順ごとにそれを行う意図を，下位の階層ではそれをどのように実現するかを記述する．したがって，各々で利用される具体的な分析手法は下位の階層で記述されることになる．このようにして分析の手順が抽象化され，複数の分野にまたがって利用可能な分析シナリオとなる．

なお，記述されたデータ分析の手順や選択肢を人間に理解できる形で管理するだけでは，結局分析者が手順をプログラミングする必要があり利便性が低いので，実行できるテンプレートの形式で保存することにした．詳細は，2.3.2 項で述べる．

2.3　分析シナリオの具体例

本節では，2.1 節で取り上げた「収益シミュレーション型」と，標準的な行動・状態から逸脱したものを検出することを目的とした「外れ値検出型」の分析シナリオについて具体例を示す．

2.3.1　分析シナリオ

分析シナリオの内容を理解するために記述した，「分析の目的」，「事例」，「訴求ポイント」の例を**図 4**，**図 5** に示す．「分析の目的」の記述の特徴であるが，各々の分析シナリオにおいてどのようなモデルを作るのかという「モデル化の対象」と，その作成したモデルの活用方法である「モデルの効果」を明確にしている点である．また，「事例」はなるべく複数あげるほうがよい．なぜならば，経験的ではあるが，顧客に提示する事例は複数あった方が，様々なことに使えそうであるという期待感を持たせることができ，顧客の業務内容のヒアリングを行う場合に，類似するような業務がないかという観点で，発想を広げる効果がある．「訴求ポイント」は，この分析を行うことによる効果を端的に記述する．これは，提示された様々な事例によって発散された発想を収れんさせる効果が期待できる．

次に，データ分析を行う手順の例を**図 6**，**図 7** に示す．「分析の目的」の場合は「モデル化の対象」と「モデルの効果」という記述であったが，それに対応づくように，「モデルの構築」と「モデルの活用」とに分けて記述している．

＜分析の目的＞
○モデル化の対象
　収益につながるアクションの種類とそのコスト，そしてそのアクションを受けた対象者がどのような効果をもたらすかを，データ分析によってモデル化する．
○モデルの効果
　ある対象者に対して行ったアクションのコストと効果の関係から期待される収益値が計算でき，その中から最も効果の高いアクションとその対象者を選択する．

＜事例＞
○購買情報
　小売業におけるダイレクトメールの送付による期待収益シミュレーション
○健康指導
　高罹患リスク者の健診値の改善対策に対する医療費削減効果の推計と，医療費削減効果のある健診値の検証

＜訴求ポイント＞
　取り得る様々なアクションに対してそれぞれの期待収益を試算し，それらを比較することで，適切なアクションの選択ができる．

図 4　収益シミュレーション型シナリオの概要
Fig. 4　Outline of Simulating Expected Profits and Losses scenario.

また，各々の手順で利用する可能性のある様々な手法の具体名を記述した例を**図 8**，**図 9** にそれぞれ示す．ここでは，分析手法が網羅的に記述されている．

2.3.2　分析シナリオの実行環境

収益シミュレーション型シナリオのテンプレートを市販ツールである「Visual Mining Studio[7),⋆1]」で作成した例を**図 10** に示す．それぞれのアイコンがデータ，もしくは分析手法を実装したモジュールに相当する．ここで，各々のアイコンは階層的になっており，たと

⋆1「Visual Mining Studio」は，株式会社数理システムの登録商標である．

＜分析の目的＞
○モデル化の対象
　事前知識による異常パターンの定義が困難な場合に，発生頻度の高いものを標準，そうでないものを外れとみなし，発生頻度の高い事象をモデル化する．
○モデルの効果
　大量のデータから異常なパターンを自動かつリアルタイムに検出し，アラーティングにより早期の危機対応を促す．
＜事例＞
○原産地表示の不正申告
　原産地表示における原産国と品目の異常な組合せの検出
○処方ミスの発見
　医療分野における疾病と処方の異常な組合せの検出
○不正アクセスの検出
　不正侵入検知システムにおける異常なアクセスパターンの検出
＜訴求ポイント＞
　異常パターンを自動検出することによって，人手を介して大量データをチェックする必要が無くなる．リアルタイム検出により，意思決定者に早期の危機対応を促すことができる．

図 5　外れ値検出型シナリオの概要
Fig. 5　Outline of Outlier Detection scenario.

＜モデルの構築＞
　回帰分析などの予測式を作成する手法を利用して，収益の要因となるユーザ属性とアクションに対応した収益の予測式を構築する．
＜モデルの活用＞
　様々なアクションを実施した後の売上の予測値とそのアクションのコストを算出する，期待収益値のシミュレーションを行う．そして，その中から最も効果の高い（期待収益値が高い）アクションを選択する．

図 6　収益シミュレーション型シナリオの分析手順
Fig. 6　Data analysis processes of Simulating Expected Profits and Losses scenario.

＜モデルの構築＞
　蓄積された申告や処方データから発生頻度の高い原産国と品目や，疾病と処方といった組合せを，標準パターンとしてモデル化する．
　ここで，標準パターンと異常パターンの識別基準となる発生頻度のしきい値は，実験的に決定する．
＜モデルの活用＞
　標準パターンと異常パターンに分類された正解データを用いて，識別基準を変動させた場合の誤識別率と期待損失の比較を行う．
　これにより，自動検出による効率化と検出漏れのリスクを定量的に評価することができる．

図 7　外れ値検出型シナリオの分析手順
Fig. 7　Data analysis processes of Outlier Detection scenario.

えば「収益予測式の構築」のアイコンでは，下の階層において回帰分析といった具体的な分析手法を用いるフローが実装されいてる．

また，統計解析的な手法で作成した予測式などは，パラメータの調整の仕方次第で過度にデータに適合してしまうことがあり[13]，これを避けるために予測式の構築で用いなかったデータで検証を行うことが重要である．そのため，モデル構築用のデータとは別に検証用のデータを用意し，検証を行うフローが加えられている．

そのほか，グラフ描画や表などの報告資料に利用する，分析結果を作成するパーツは，モデルを活用した効果を示すためのノウハウであるため，テンプレートに入れておくことにした．このような対策により，抜け漏れのない処理フローの構築や，分析報告書の高品質化が可能となると考えられる．

2.4　分析シナリオの利用方法

今回作成された分析シナリオであるが，データ分析担当者内でのノウハウの共有に加え，顧客とのヒアリングの際の初期資料として利用することができる．データ分析のトライア

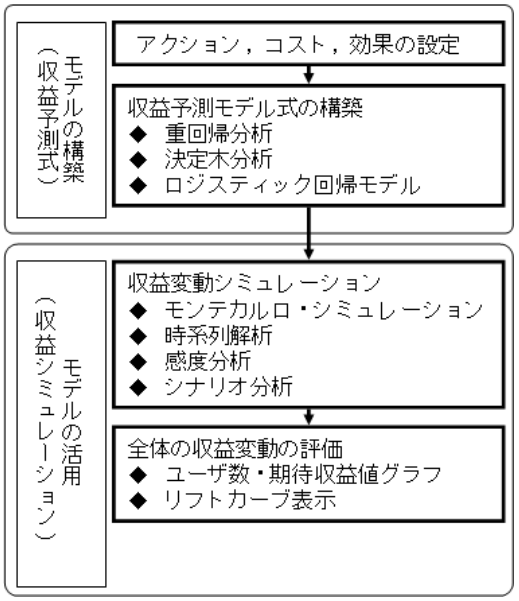


図 8　収益シミュレーション型シナリオに適用可能な分析手法

Fig. 8　Data analysis processes and alternative methods for Simulating Expected Profits and Losses scenario.

ルを行う際の大まかな流れは，提案フェーズ，分析フェーズ，レポーティング・フェーズに分かれるが，初期の提案フェーズでは顧客から具体的な課題を引き出すことが重要である．この場面において，顧客の中でデータ分析によって解決される業務課題が明確であることはまれであり，我々が類型化した分析目的を示しながら議論することで，お互いに具体的な課題をイメージしながら業務の詳細をヒアリングすることができるという効果がある．

このようなやりとりから，実際の業務で求められるデータ分析がどの分析シナリオに相当するかを把握し，それに沿った提案資料を作成することで，ニーズにあったデータ分析内容を決定することができる．

3. 分析フレームワークの効果検証実験

3.1 実験の概要

本節では，分析シナリオを用いて，過去の実績がない業務分野のデータ分析課題に取り組む場面を想定し，我々の提案の効果および活用できる条件を実験により検証する．具体的には，データ分析の業務経験がない被験者をチームに分け，分析シナリオを利用するチームと利用しないチームのそれぞれで，データ分析の課題を解く作業を行うこととした．今回利用

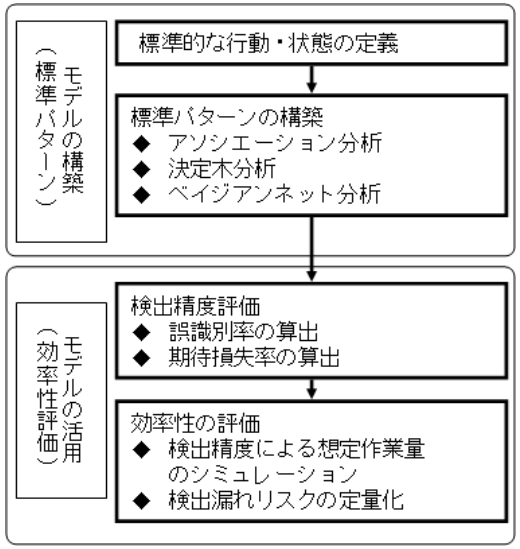


図 9　外れ値検出型シナリオに適用可能な分析手法

Fig. 9　Data analysis processes and alternative methods for Outlier Detection scenario.

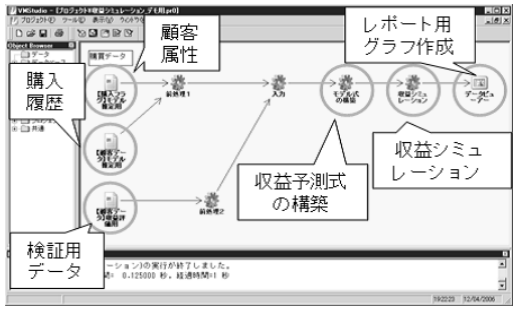


図 10　テンプレートの例

Fig. 10　An example of a template.

した課題は 2 つである．課題 1 は，2.1 節で例にあげた「収益シミュレーション型」を健康保険データの分析に適用する例である．課題 2 は，仕入れた材料の原産地に不正がないかを管理する業務をイメージした，過去の申告データからよく見られる内容を標準パターンとしてモデル化し，標準パターンから外れた申告を検出してアラートをあげる，「外れ値検出型」の分析である．なお，それぞれ用いたデータの量は課題 1 が項目数が 20 でデータ数は 20,973，課題 2 はそれぞれ 5 と 2,000 である．

ここで，課題 1 で分析シナリオを利用したチームは課題 2 では利用せずに取り組むこととした．もう一方のチームは逆に，課題 1 で分析シナリオを利用せず，課題 2 では利用することとした．これは，課題 1 が業務を初めて経験する場面，課題 2 が業務を経験した後の場面を想定しており，提案する分析シナリオが，ノウハウの共有という面においてどのような条件で効果を発揮するかを検証するために行う．

また，効果の測定は主に 3 つの観点で行う．データ分析業務は，顧客の課題をヒアリングし分析の目的や手順を決定する提案フェーズ，それをもとにデータ分析プログラムを構築する分析フェーズ，その結果を報告するレポーティング・フェーズがある．それぞれのフェーズでの提出物は，提案フェーズでは分析手順書，レポーティング・フェーズでは分析報告書であり，それぞれドキュメントの質と作成にかかった作業時間を評価する．また，テンプレートは分析フェーズの短縮に効果があると考えられるため，この作業時間を計測する．

提案フェーズの評価では，分析手順書に記載された 1) 前処理，2) モデル構築，3) モデル検証，4) 結果評価の 4 つの手順に対して，i) データの記載内容と，ii) 手順と手法の詳細さの 2 つの観点で，データ分析業務のエキスパートが，それぞれ 5 段階評価の 8 項目で計 40 点満点の得点化を行った．

なお，被験者にはどの分析シナリオであるかは教えずに，実際の業務の提案フェーズと同様に分析手順書を作成した後に報告を行うことにした．ここで提案書の内容を確認し，課題に対する適切な分析提案ができるよう，指導することを繰り返した．この結果として，最終的にはどのチームも想定する分析シナリオとなった．ここでは，ドキュメントを記述する作業時間の合計を計測の対象とした．

レポーティング・フェーズの評価では，分析報告書に記載された，1) 分析に用いたデータ項目の記載，2) それを表現する表やグラフ，3) 結果を導き出すために行った分析の手順，4) 考察，および 5) データから分かった特徴的な事柄について，データ分析業務のエキスパートが，それぞれ 5 段階評価の 5 項目で計 25 点満点の得点化を行った．また，分析プログラム構築，およびドキュメント作成のそれぞれに要した時間を計測した．なお，データの成形を行う前処理に関しては，事例ごとに対応方法が異なることが多く，テンプレート化の対象としていないため，今回は評価の対象外とした．

被験者は 2 人 1 チームで，A，B，C，D の計 4 チームとした．被験者のデータ分析のリテラシ・レベルを確認するために，前処理などで使うデータベースや標準的なプログラミング言語と，データ分析の基礎となる統計解析用の言語やツールについてテストを行った．この結果をもとに各チームのリテラシ・レベルが均等になるように組合せを決定した．具体的には，チームメンバで各々のリテラシを補完しあえるように，各チーム内で各々のリテラシの高い方の得点を合計した値を算出し，この値のチーム間のばらつきが最小となるように組合せを決定した．これらのチームに対して，課題 1 は C チームと D チームが分析シナリオありで，A チームと B チームが分析シナリオなしで分析を行った．課題 2 はその逆である．また，分析シナリオを使用するチームと使用しないチームは，それぞれ別室で作業を行うこととし，分析プログラムの構築はすべてのチームで同じ市販ツールを利用することとした．なお，「与信管理型」を想定した課題を例題にし，被験者に作成してもらいたい分析手順書と分析報告書のイメージや，利用する市販ツールと想定する分析手法に対して事前にレクチャを行い，模擬データによる自習の時間をとることにした．

以上の評価基準をもとに，業務経験がない場面での効果と業務経験がある場面での分析シナリオの効果をそれぞれ検証する．

3.2　実験の結果

3.2.1　データ分析経験がない場面での効果

表 3 に，課題 1 に対して分析シナリオがある場合とない場合で，分析手順書，分析報告書のそれぞれ質の評価結果，分析手順書，分析プログラム，分析報告書のそれぞれの作業時間，および一元配置の分散分析[14] を用いて，分析シナリオの有無による違いを 5%の有意水準で検定した結果を示す．リテラシ・レベルとは，リテラシ・チェックテストにおける各チームの順位に対応する．検定結果が「有意」とは分析シナリオの有無による違いがあることを示しており，「無意」とはその逆である．また，順位相関とは，被験者のリテラシ・レベルとドキュメントの質や作業時間との関連性が考えられるため，分析シナリオの使用有無とは無関係に，チーム間のリテラシの順位とそれぞれの評価結果の順位に対してケンドールの順位相関係数[11] を算出した結果である．

この結果から全般的に分析シナリオの有無による違いはなく，分析手順書の質や作業時間はリテラシ・レベルとの相関が 0.67 とやや高いため，分析シナリオの効果があるとはいいがたい．ただし，分析報告書の質やプログラム作業時間に関しては，順位相関が負の値を

表 3　課題 1 の質の評価と作業時間（分）の結果
Table 3　Results of quality score of documents and time[m] in the 1st experiment.

分析	リテラシ・	分析手順書		分析報告書		プログラム
シナリオ	レベル	質	作業時間 (m)	質	作業時間 (m)	作業時間 (m)
あり	2	10	95	25	570	30
	4	8	110	23	590	50
なし	3	17	90	20	500	460
	1	18	65	15	240	1,020
検定結果		有意	無意	無意	無意	無意
順位相関		0.67	0.67	−0.33	0.67	−0.33

表 4　課題 2 の質の評価と作業時間（分）の結果
Table 4　Results of quality score of documents and time[m] in the 2nd experiment.

分析	リテラシ・	分析手順書		分析報告書		プログラム
シナリオ	レベル	質	作業時間 (m)	質	作業時間 (m)	作業時間 (m)
あり	3	34	140	24	660	140
	1	38	60	23	100	300
なし	2	20	150	13	480	1,320
	4	19	155	15	380	1,330
検定結果		有意	無意	有意	無意	有意
順位相関		0.67	0.67	0.00	0.33	0.33

とっていることから，リテラシ・レベルの影響よりも分析シナリオの効果が期待できる．

以上から，データ分析経験がない場合において，分析報告書やプログラムの作業において分析シナリオの効果は期待されるものの，全般的に効果が薄いという結果になった．

3.2.2　データ分析経験がある場面での効果

課題 1 と同様に，課題 2 の結果を**表 4** に示す．この場合は，分析シナリオを用いることで，分析手順書，分析報告書ともに質の向上が見られるという結果になった．一方で，作業時間に関しては分析手順書や分析報告書といったドキュメントを作成する時間の短縮の効果は見えないが，プログラムを作成する時間については効果があるという結果となった．順位相関に関しては，分析手順書の作業において 0.67 とやや高くリテラシ・レベルの影響を否定できないものの，分析報告書やプログラムの作業においては 0.00〜0.33 の間と全般的に低い値であることから，作業全体を通して分析シナリオの効果があるといえる．

ここで，プログラム作成時間の短縮効果が示されたが，これは分析シナリオのテンプレートを利用することで，テンプレートをカスタマイズする程度の作業でプログラムの作成ができたからである．このことから，過去の事例を蓄積するだけではなく，分析シナリオとして分析内容をドキュメント化することで，テンプレートの効果がより向上することを示唆している．

3.3　実験結果の考察

課題 2 の分析シナリオありのチームは，課題 2 において分析シナリオを初めて利用したにもかかわらずドキュメントの質の向上が見られたことから，データ分析の作業に慣れることが分析シナリオの理解に影響すると考えられる．

一方で，ドキュメント作成時間の短縮は見られなかったが，分析シナリオの中から対応するものを選択するために，内容を理解することに時間がかかったという被験者の意見があった．また，分析報告書の作成時間の比較であるが，分析シナリオなしのグループではテンプレートが用意されていないため，分析結果の出力（グラフ，表）が少なく，結果の比較などの推敲の材料が少ないため，結果的に作業時間も少なくなっていた．反対に，分析シナリオありのチームでは，分析プログラムを作成する時間が短縮できた影響で気持ちに余裕ができ，分析報告書の作成により時間をかけることができたという意見が見られた．以上のことから，ドキュメント作成作業においては，時間短縮の効果よりも質の向上に効果があることが分かった．

実験終了後に収集したアンケートでは，分析シナリオで使用されている用語になじみがなく理解に時間がかかってしまったため，分析手順を設計する実質的な時間が短くなってしまったという意見と，事例が記述されていたため内容の理解が深まったという意見があった．特に理解に時間がかかった点として，汎用的な表現で記述されているため，分析シナリオ間の違いが分かりにくく，事例も含めて分析シナリオ全体の理解が必要であったことが指摘された．

1 章で述べたように，業務分野を特定して事例を蓄積したり，分析手順をテンプレート化したりする方法の場合は，業務分野やその事例に特化した表現がされており，その業務分野に対する業務知識があれば，分析手順を理解することは容易であるが，他の業務分野に展開することを考えると，個々の事例から本質的な部分をくみ取って，別の業務分野に展開する高度なスキルが要求され，実質的には困難になる．一方で，分析シナリオを用いる場合は，業務分野を特定されず汎用的な表現をとっているため適用範囲が広いが，分析シナリオを選択する場面において，個々の分析シナリオの狙いなどが抽象化されて記述されているため分かりにくく，すべての分析シナリオをあらかじめ理解しておく必要があったといえる．この分析シナリオを選定する問題に対しては，分析シナリオの違いを説明した分析シナリオ

選定のためのガイドを用意することによって，より効率的に分析シナリオが活用できるようになると考えられる．

今回の検証における分析シナリオの効果をまとめると，データ分析の経験があるという条件では，特に効果のあった分析報告書やプログラムを構築する分析フェーズにおいて，課題2の結果から，分析シナリオを利用することで分析報告書の品質で平均で約67%の向上，分析報告書，プログラム作成を合わせた作業時間を平均で約34%に短縮できたといえる．また，プログラム作成の時間だけでは平均で約16%に短縮可能である．

4. まとめと課題

データ分析作業で得られるノウハウを蓄積するために，データ分析の課題を目的別に類型化し，それぞれをシナリオ化するためのフレームワークを提案し，実験により分析シナリオが有効に機能する条件を示した．実験の結果から，データ分析作業の経験があるという条件では，質の高い分析手順書や分析報告書の作成が可能であり，全体的な作業が短縮できることを示した．以上のことから，我々の提案する分析シナリオが，ノウハウの蓄積および共有という目的を達成していると考えられる．

今回は過去にたずさわった約100件の事例をもとにサブシナリオを含めて13の分析シナリオにまとめたが，これは過去の事例をまとめたものにすぎず，データ分析業務のノウハウ共有の取組みの第1段階だと考えている．したがって，今後蓄積されていく事例に対して，まったく新しい分析シナリオを追加する場合や，サブシナリオとして階層的に追加する場合など，様々な条件にあった類型化方法を確立することが課題である．

なお，業務分野ごとの個別性が高いと考えられる，分析対象項目の選択，外れ値除去，データ成型方法，業務知識の継承といった，対応できていないノウハウについてもさらなる検討が必要である．

謝辞　本研究の実験に多大なるご協力をいただいた，早稲田大学創造理工学部大野高裕教授，後藤允助手，および大野研究室の学生の皆様，本研究の機会を与えていただいた当社技術開発本部ビジネスインテリジェンス推進センタ上島康司部長，松永務課長，安藤陽介課長ならびに，日頃議論していただいている同僚諸氏に感謝します．

参　考　文　献

1) 武田浩一：ビジネス・インテリジェンスと人工知能技術，情報処理，Vol.47, pp.723–728 (2006).
2) Codd, E.F., C.S.B. and Sally, C.T.: *Providing OLAP (on-line Analytical Processing) to User-analysts: An IT Mandate*, Codd and Date, Inc. (1993).
3) Berry, M.J.A. and Linoff, G.S.: *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 2nd Edition, Wiley (2004).
4) 元田　浩，鷲尾　隆：機械学習とデータマイニング，人工知能学会誌，Vol.12, pp.505–512 (1997).
5) IBMビジネスコンサルティングサービス：実践!「経営の見える化」プロジェクト，日経BP社 (2006).
6) Witten, I.H. and Frank, E.: *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann (2000).
7) 株式会社数理システム．http://www.msi.co.jp/vmstudio/
8) エス・ピー・エス・エス株式会社．http://www.spss.co.jp/clementine
9) SAS InstituteJapan株式会社．http://www.sas.com/japan/
10) 秋葉さくら：意志決定と収益に貢献するBIテンプレート，*INTEC Technical Journal*, Vol.3, pp.23–33 (2004).
11) 柳井晴夫，高根芳雄：新版多変量解析法，朝倉書店 (1977).
12) Quinlan, J.R.: *Induction of decision trees*, Springer (1985).
13) 坂野　鋭，山田敬嗣：怪奇!!次元の呪い—識別問題，パターン認識，データマイニングの初心者のために—，情報処理，Vol.43, pp.562–567 (2002).
14) 永田　靖：入門統計解析法，日科技連 (1992).

(平成20年3月20日受付)
(平成20年7月 9 日採録)

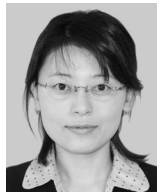
(担当編集委員　鬼塚 真)

末永　高志

平成9年早稲田大学理工学部経営システム工学科卒業．平成11年同大学大学院理工学研究科修士課程修了．同年株式会社NTTデータ入社．パターン認識，データ分析技術の実用化研究に従事．

山中　啓之

平成 10 年大阪大学大学院基礎工学研究科情報数理系専攻修士課程修了．同年株式会社 NTT データ入社．データ分析技術の実用化研究に従事．



高橋　彰子

平成 11 年慶應義塾大学大学院理工学研究科管理工学専攻修士課程修了．同年日本電信電話株式会社に入社し，通信トラヒック解析等の研究に従事．平成 15 年株式会社 NTT データに転籍し，データ分析技術の実用化研究に従事．

東　　陽子

平成 13 年東京大学大学院理学研究科修士課程修了．同年株式会社 NTT データ入社．遺伝子データ解析の研究開発を経て，平成 17 年より分析統合フレームワーク開発に従事．

佐治　美歩

平成 14 年大阪大学大学院基礎工学研究科情報科学専攻修士課程修了．同年株式会社 NTT データに入社．以来，データ分析技術の実用化研究に従事．

矢野　順子

平成 15 年東京理科大学大学院工学研究科経営工学専攻修士課程修了．同年株式会社 NTT データに入社．以来，データ分析技術の実用化研究に従事．

中川慶一郎

平成 4 年早稲田大学大学院理工学研究科修士課程修了．同年株式会社 NTT データ入社．早稲田大学大学院理工学研究科博士課程満期退学．株式会社 NTT データ技術開発本部主任研究員．博士（工学）．

関根　　純（正会員）

昭和 57 年東京大学大学院工学研究科計数工学専攻修士課程修了．同年日本電信電話公社（現 NTT 入社）．平成 17 年より株式会社 NTT データ技術開発本部副本部長．博士（工学）．