人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第19回)
SIG-AM-19-04

# データの統合化と視覚化によるデータ分析統合ツールPADOCの提案

# Proposal for Data Analysis Tool PADOC by Data Integration and Visualization

中井 眞人[1*]　角田 善彦[1]　林 久志[1]　村越 英樹[1]
Masato NAKAI[1]　Yosihiko TSUNODA[1]　[1] Hisashi HAYASHI　Hideki MURAKOSHI[1]

[1] 産業技術大学院大学 産業技術研究科
[1] School of Industrial Technology, Advanced Institute of Industrial Technology

**Abstract:** In Analects, there is a saying "visiting old, learn new". This means to investigate old things, in order to obtain new knowledge and insights.The process of data analysis could be interpreted as an act to analyze the data of the past, discover new knowledge and insights mathematically and make good use of them toward better future. Until end of the 20th century data was very valuable and less reliable, but now the accumulation of data became remarkable due to the explosive spread of the Internet society in recent years. However, proper use of data has not yet been established. The reason for this is that since the data is accumulated according to the operation of each business, there is no standardized analytical method because the accumulation state of data varies. Therefore, it is necessary to edit and integrate data by processing on computer for analytical purpose. Furthermore, it is necessary to examine whether the edited data is appropriate for analysis. To do so, it is convenient to have a tool that analyzes data while visually showing and checking data by editing and examining data. This paper proposes a graphical analysis integration environment "PADOC" to facilitate data editing and data review.

## 1　はじめに

論語の「温故知新」は昔の事を調べて，そこから新しい知識や知見を得ること（大辞林）[1] であるが，データ分析は過去のデータを分析して，その知見を数理的に発見することともいえる．20 世紀の後半まではデータは分析目的に合わせて収集することが多く，データ量も少なく信頼性も低いため，結果を出しても検証が煩雑であった。しかし近年のネット社会の爆発的な広がりでデータの蓄積は著しいものになったが，未だにデータ分析の適切な方法が確立されていない．これはデータが個々の業務の運用に合わせて蓄積され，データが多種多様に存在する一方，分析に適したデータが直接見つかることは稀で，見つかっても分散されている場合が多いからである．現在はデータ量は多くなったが，分析したいデータを作成するにはデータを適切に加工し統合することが必要になっている．このデータの加工と統合は前処理と云われ，一般的には全工程の 7 割が前処理に費やすと云われている．

*連絡先：産業技術大学院大学 産業技術研究科 創造技術学科
〒 140-0011 東京都品川区東大井１丁目１０－４０
E-mail:b1617mn@aiit.ac.jp

しかし近年急速に普及した無償の分析ツール $R$ や Python は数理モデル構築に重点を置いており，データ編集が容易でないことが多い．そのため $R$ や Python の前処理に特化した詳細な解説本「前処理大全」[2] が最近出版されている．

本稿はデータ編集とデータ分析を容易にするためのグラフィカルな分析統合環境 PADOC[1](Process Analysis by Data Oriented Composition) を提案する．これはデータ編集にはコマンド環境を提供し分析には分析過程をイメージし易い様にグラフィカル環境の両方を提供している．図 1 の左図は提案ツールのデータ編集のコマンド記述の一部を選択して実行している例で，右図はグラフ上のアイコンを繋いでデータ分析過程を構築している例である．

本稿ではデータ分析の目的を全体像の把握，比較検討，仮説検証，知識発見に分ける．全体像の把握に要するデータ編集では「前処理大全」にある Python と提案ツールの記述との比較を行ないその簡潔性を示す．比較検討や仮説検証及び知識発見では提案ツールが十分な機能を提供しグラフィカルな環境が効果的であることを示す．

図 1: Left:command mode Right:graphical mode

## 2　先行データ分析ツール

データ分析ツールは古くから有償の SAS[3]，SPSS[4]，S-PLUS[5]，Matlab[6] があり，近年では無償の $R$ や Python が広く使われる様になっている．SAS は 1976 年から統計パッケージとして販売され最も実績がある．その実績から米国の医薬系の申請では SAS での報告が求められてきた [7]．SAS はデータが貴重で計算機が貧弱な時代に出現したので，結果の検定は得意としているが，高性能な計算が必要なグラフィカルな表示や直近のアルゴリズムの実装が貧弱である．逆に貧弱な計算機でも稼動する様にデータはテーブル形式を対象にしており，データ加工は一行単位に同じプログラムを適応するだけなので手続きが明瞭で記述し易い．関数型の S-PLUS は現在の無償の $R$ として発展した．Matlab は行列演算を得意とし豊富な科学計算ライブラリィを提供したが，無償の Python も殆ど同等の機能が提供される様になり差別化が困難になってきた．$R$ は初めて無償で本格的な統計パッケージであり広範に広まったが，関数型の記述や大規模なデータを不得意としているので Python に比べて劣勢にある．

一方隆盛を見せている無償の Python は高速な計算機とメモリーを贅沢に使ってプログラム言語の宣言やメモリー管理を不要にしてロジックが組みやすい記述を提供している．また無償化によって最新の深層学習系のロジックや豊富なグラフィック表現が次々提供されて加速度的に発展している．しかしオブジェクト指向型プログラム言語なのでオブジェクトに依存した多様なメソッドを駆使しなければならず習熟が難しい．またビックデータの前処理に既存のＰＣを連結して分散処理する無償の Hadoop[8] がある．これはクラウドの様な刻々と大量に流入するデータの選別やデータの変換には威力があるがデータ整形だけで分析機能がない．

提案ツール PADOC はコマンドベースではデータ加工を SAS の様にレコード単位に記述する平易な表現を採用し，一方グラフィカルな視覚表示では分析過程や分析結果を評価し易い環境を提供している．

[1]windows 7 8 10 で稼動，Python インストール要

## 3　提案ツールの説明手順

近年ではデータ分析の精度を争うサイト Kaggle[9] が出現し，データ分析技術の向上に大きく寄与している．この分野での知見が広がるのは望ましいが，データ分析とは理論や技術を駆使して分析精度を競うものと認識される傾向がある．しかしこれにはデータの前処理が抜けており，結果も精度を競うだけで実務上大事なモデルの頑健性や結果への説明力が抜けている．

実務用にデータ分析を定義したものとして総務省の「高度 ICT 利活用人材育成プログラム開発事業（実践偏)」[10] の資料がある．この資料ではデータ分析の用途を次のように分類している．本稿ではこの項番に沿って提案ツールの機能を説明する．

4. 全体像の把握
5. 比較検討
6. 仮説検証
7. 知識発見

4. 全体像の把握については，提案ツールのデータ編集の容易性を示し，5. 比較検証　6. 仮説検定　7. 知識発見では提案ツールの提供モデルとグラフィカルな表示環境が十分な機能を提供していることを示す．

## 4　全体像の把握

分散されたデータでは全体像を把握し難いので，一般的にデータを編集して統合する前処理を行う．しかし統合するとデータの定義は拠り所を失うので，データ定義はシステム運用に従って確認を行う必要がある．データの誤解釈は後続する分析を無駄にしてしまうので極力避けなければならない．

### 4.1　全体像の把握ツール

全体像の把握は各データ項目の充足状態や分布状態から分析に耐えられるか見る場合が多い．提案ツールでは分析対象項目とその他の項目との関係の強さ順に充足状態や分布も表示するツール [11] が提供されている．図 2 の例はローン破綻と関係が高い項目のランキング表示で，持ち家状態 (home)，ローン金額 (amount)，貸出し期間 (mon) が高い順になっていて，各項目の分布状態も示されている．持ち家状態 (home) の分布では賃貸や借家などの流動性が高い先のローン破綻率が高いことが示されている．

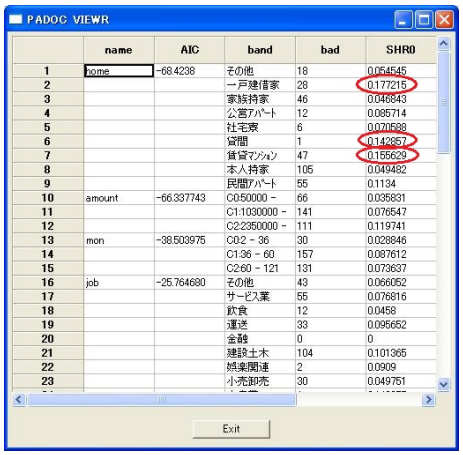


図 2: Relation ranking for loan collapse by AIC

## 4.2 データの前処理

データの前処理に関しては「前処理大全」の項目に沿って提案ツールの優位性を述べる．前述した「前処理大全」の前処理では次のセクションで記述されている．

(1) 抽出 (2) 集約 (3) 結合 (4) 分割 (5) 生成 (5) 展開

以下この手順に沿って提案ツールの前処理の優位性を述べる．

### 4.2.1 データ抽出（抽出）

データは業務運用のために蓄積されているので，分析用に必要な項目を抽出する必要がある．下記に示す様に欠損を除いて項目を抽出するには Python はメソッド関数を駆使するが，提案ルールは項目の列挙と抽出条件だけなので簡潔である．

- Python

```
#項目選択メソッドの使用
select_tb = bankr.loc \
[:['home','amount','job',]]
#欠損用の削除メソッドの使用
select_tb['amount'].dropna()
```

- 提案ツール PADOC

```
/* データ呼出し */
get bankr.csv@;
/* 項目選択 */
select home amount job;
/* 欠損レコードの削除 */
if(amount == ?) delrec;
```

### 4.2.2 サマリー処理 (集約)

一般にデータ分析では各レコードが独立であることが前提であるが，次の様な理由で独立を損なわれ結果に歪みが生じてしまう．そのためサマリー処理が必要である．

1. データが重複していると，重複が多いレコード寄りの結果となってしまう．

   例えば顧客別の明細に多寡がある場合，明細の多い顧客の特性が結果に反映されてしまう．この場合は顧客毎に明細をサマリーする必要がある．

2. 分析期間の長短によって結果が異なる．

   分析期間が異なると外的要因に晒されている期間が異なるので同じ条件のデータにならない．この様な項目は期間平均に直す必要がある．

以下は米国の業務種別（jobatnm)，人種別 (minoritynm) の給与 (salnow) をサマリーした例である．Python はメソッド関数を連結しているが，提案ツールでは，sumup コマンドを使って簡潔な表現でサマリー処理ができる．

- Python

```
reult = bankr.groupby \
(['jobcat','minority']) \
['salnow'].sum().reset_idex()
```

- 提案ツール PADOC

```
/* データ呼出し */
get bankr.csv@;
/* jobcat(職種) と minority(人種) 別にサマリー */
sumup  salnow by jobcat minority
```

### 4.2.3 データ結合 (結合)

データを統合するには分散データのＩＤを介して連結する必要がある．一般的には分散データ上で不要な項目を削除してから統合するのが普通である．この様な統合ではデータテーブル間の関係が容易にイメージできる必要がある．提案ツールではこの目的のためコマンドベースと図 3 の様なグラフィカル環境を設けている．図 3 は各分散データから項目を抽出して逐次合成する過程をアイコンで連結して実現している．一般に Python や「R」はオブジェクト指向のメソッド関数

を使った複雑な記述になり，全体的な統合過程をイメージするのが難しい．

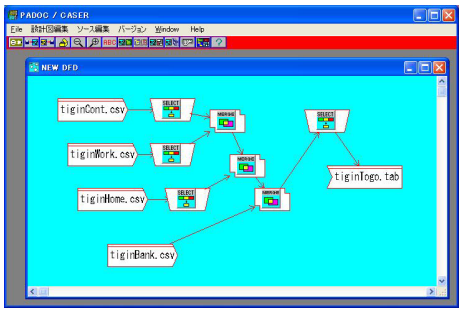


図 3: Data flow graph for Integration

#### 4.2.4 データの分割（分割）

分散データを統合したデータをマスターデータと云うが，一般には以下の理由でデータ分割する場合が多い．

1. 分割による欠損の排除

   一般に統合すると業務上の理由で多くの項目で欠損が存在する．例えば株式会社であれば財務データは存在するが，個人営業会社では得られない場合が多い．この場合は株式会社と個人営業会社に分割して分析すると欠損を回避できる．

2. 分析の過学習を回避

   データ偏在を避けるためデータをランダムに分割し混合する交差検証でモデルの安定性を図る．

ランダムにデータを分割する場合は，提案ツールでは明示的に乱数の範囲を指定する

- Python

```
row_no = list(range(len(bankr)))
#4 分割指定
k_fold = KFold(n_splits=4,shuffle=True)
#4 分割
for train_cv_no in k_fold.split(row_no) :
  bank = train_data.iloc[train_cv_no,:]
```

- 提案ツール PADOC

```
get bankr.csv@;    /* データ呼出し */
rnd = random;      /* 一様乱数付与 */
/* 4 分割 */
     if(rnd <= 1/4) outrec bank1;
     else if(rnd <= 2/4) outrec bank2;
     else if(rnd <= 3/4) outrec bank3;
     else                outrec bank4;
```

### 4.3 データ加工（生成 展開)

データ加工の目的は分析に合う様に項目を作り出すことである．業務運用上蓄積されているデータだけでは，分析目的に合う項目が存在するとは限らない．分析用のデータを新たに生成する場合がある．

1. 市場データ等の公開若しくは有償で入手できるもの

   倒産の予測 [12] では日銀の市場データ [13] と自社の倒産推移とで図 4 の様に金利が倒産の推移に一年先行している事を使って 1 年後の倒産予測をしている．

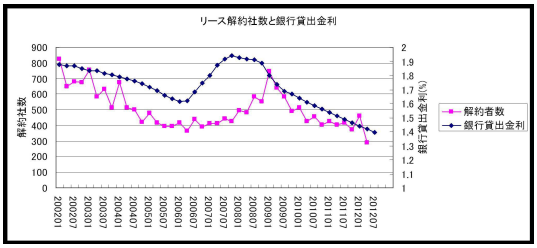


図 4: Relation of bankruptcy and interest rate

2. 自己資本比率の様に項目間で計算できるもの

   自己資本比率 = (総資本 - 負債)/総資本

3. 外れ値の補正やデータ値の偏在を避けるため対数化や正規化する

   一般に値段等の正の値を持つものは対数化すると正規分布になることが知られている．下記は対数化した例であるが，Python はオブジェクト志向型の言語なので，オブジェクト毎のメソッド関数を駆使する必要がある．一方提案ツールは平易な表現で全レコード対数化できる．

   - Python

```
reserve_tb['total_price_log'] = \
reserve_tb['total_price']. \
apply(lambda x:np.log(x/1000+1))
```

   - 提案ツール PADOC

```
get reserve_tb; /* データ呼出し */
togal_price_log=log(total_price/1000+1);
```

# 5　比較検討

本節以降の比較検討，仮説検討，知識発見について提案ツールの分析モデルは，グラフィカルな表現によって十分な性能を提供していることを示す．

一般に比較検討は区分による相違を見ることが多い．図 5 の右図のろうそく図は例は職種別の給与の分布を示し，一般職（左 3 業種）と資格職（右 4 業種) とで大きな相違が見られる．

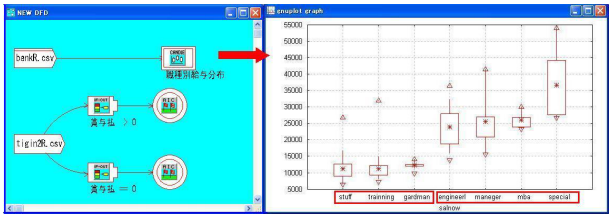

図 5: Candle chart for salay by job category

上図 5 の左図の分割プロセス図に示す様にローン返済で賞与払いの有無でデータ分割して，ローン破綻に関してランキング表示すると全項目について一度に比較することができる。図 6 の左図は賞与払い先で一定の賞与が見込まれる従業員のデータと見られ，右図は賞与払いが難しい経営者のデータと考えられる．ローン破綻が一番強く関係するのは従業員は借入金額 (amount) の多寡で，経営者は住宅の所有状態 (home)（即ち居住流動性）と相違している．

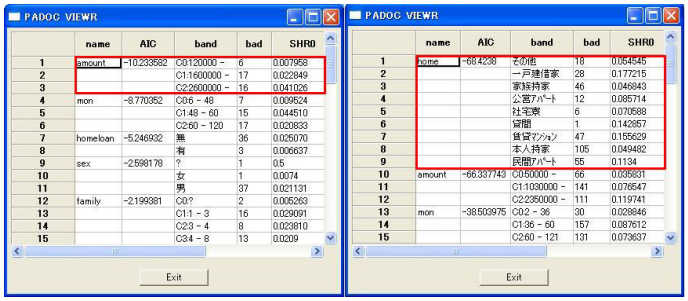

図 6: AIC ranking. Left:bonus Right:non bonus

# 6　仮説検証

仮説検証は，分析対象項目を他の項目で予測する仮説モデルの信頼性を検証する場合が多い．このモデルは分析対象データに合う様に予測するので一般には教師付モデルと云う．検証は予測値と分析対象の実測値の差を示す精度指標で行う．

## 6.1　教師付モデル

1. 分析対象が 2 値ならロジステック回帰モデルを使う [15]

   しかし図 7 の左図の様に全く線形的な分離が難い場合は SVM[14] が適切である．提案ツール SVM は右図の様にデータを高次元に写像してから線形分離している．

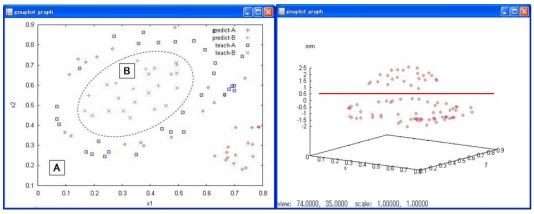

図 7: Impossible classification zone by linear

2. 分析対象が実数値なら重回帰モデル

   一般に重回帰は決定係数で検証する．図 8 の左図は 3D 図での回帰結果である．3D では回帰値は網掛けの平面になり、この平面上に殆どの点が載っていることがわかる．

3. 分析対象が時間経で劣化するならハザードモデル [16]

   図 8 の右図の例は格付別の会社の生存件数の経過予測したもので，下位格付ほど生存件数の降下が大きいことを示すことができている．

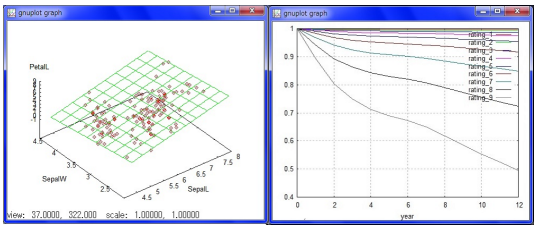

図 8:　Left 3D regression Right:cox hazard for bankruptcy

4. 時系列の推移予測で定常性があるなら ARIMA モデル，非定常性ならカルマンフィルターが使われる [25]

   提案ツールでは図 9 の左図の様にアイコンを繋ぐことでデータ加工過程を編集することができる．この図は株価をトレンド成分とフーリエで低周波を除去して，定常波にしてから ARIMA モデルで予測するプロセスを示したものである．右図の赤線部分は予測を示している．

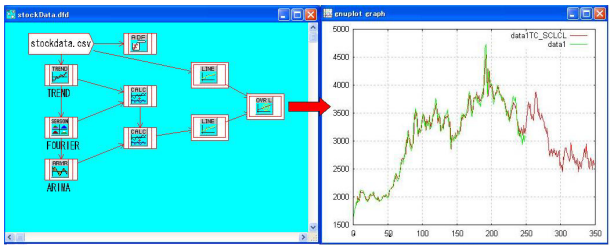


図 9: Left:analysis flow Right:prediction by ARIMA

## 6.2 精度と頑健性の問題

この様な教師付モデルは，分析対象と関係が強い項目（特徴量）の重み付線形和で表現されているので次の 3 つの観点で精度の検証が大切である．

1. 精度と頑健性はトレードオフの関係になっている．

   - 分析対象と関係の低い項目を追加しても精度は向上する．
     これはノイズでモデルを説明する過学習状態となっている．
   - レコードに重みを付与するブースティング法
     これは誤判別率に比して重みを付与して改善する方法である．

   何れも過学習すると試験用のデータでは精度は劣化する．精度と頑健性はトレードオフの関係があり，一般的には精度を高めると頑健性は劣化する場合が多い．

   分析対象が 2 値の場合では提案ツールは図 10 の左図の様に Boosting[17] を繰返して精度を上げることができ，右図の AR 曲線 [2] の様に Boosting による精度（緑線) と試験データでの精度 (赤線) を比較して，無理に精度を向上させていないか確認することができる．

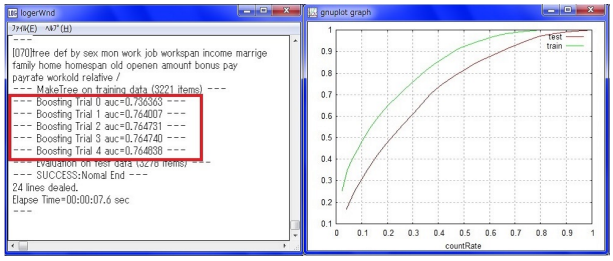


図 10: Right:boosting process Left:comparison of AR curves

[2] AR 曲線は横軸に事象の生起確率の高い順に並べ，縦軸に実際事象が起きた件数の累計をプロットして繋いだ線である．曲線の膨らみが大きい方が精度が良い．

分析対象が実数値の重回帰の場合は提案ツールでは正則化項を入れた Lasso モデルで過学習防止している．

2. 複数の分析手法を比較して優良モデルを選択する

   図 11 の左図は提案ツールでの判別木 (緑線) とロジット回帰 (赤線) の結果をグラフで重ね合わせて比較する過程を示した図である．右図は AR 曲線による精度の比較結果である．

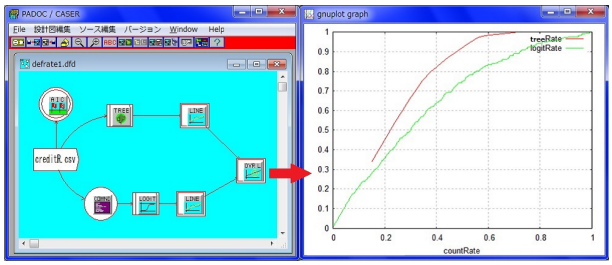


図 11: Right:process of Tree and Logit Left:comparision of AR curves

3. データが豊富にあればモデルの適合性を計るＡＩＣ [11] やＢＩＣ [19] を使わない．

   データが貴重な時代はモデルの適合指標として解釈が難しいＢＩＣやＡＩＣ等が使われたが，データが豊富な場合はデータを学習用と試験用に分割して図 10 の右図の様に精度比較をした方が合理的である．特に時間経過での頑健性を期待するなら，データを時間経過別にデータ分割して経年変化にも頑健性があるか検証すべきである．

# 7 知識発見

前節の仮説検証が目標データがある教師付モデルであれば，知識発見は教師データがなく一般的にはデータから隠れた要因を推定するモデルが主流である．知識発見では，隠れ変数推定ベイズモデル，グラフィカルモデル，最適計画問題がある．教師データが無いので結果の検証は難しい．結果は主に図やグラフで視覚的に説明される場合が多い．グラフィカルモデルはグラフ上でデータ項目間の最適な関係を示すものである．グラフに依らない最適問題として最適計画法や強化学習がある．提案ツールでは次の様なモデルを提供している．

1. 隠れ変数推定モデル

   1.1. MCMC(markov chain monte calro) [18]

隠れ変数をマルコフ連鎖でサンプリングして最大尤度を持つ値を探索する

1.2. EM 法 (expectation-maximization) [19]

隠れ変数の期待値で尤度を最大化する値を推定する

1.3. 変分ベイズ法 [19]

隠れ変数で仮定した変分の下界を最大化して値を推定する．

図 12 の例は米国の間欠泉 (Old Faithfull) の噴射間隔と噴射高のプロットしたものである．左図は EM 法　右図は変分ベイズ法で各々の集団の所属率を隠れ変数として推定している．

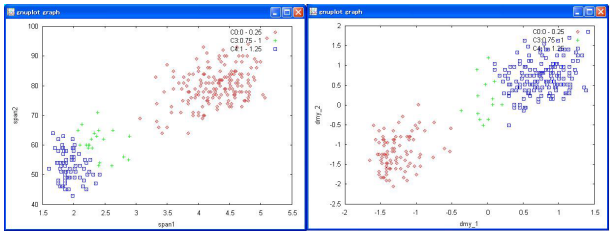

図 12: Left:EM Right:Variational Bayes for OldFaithfull

1.4. データの近傍状態を推定する樹系図

図 13 の左図の例は 10 人の成績がどの様に類似しているか示す樹系図である．赤で囲んだ Mia のみ離れているが右図の 3 D図でも外れ位置にあることがわかる．

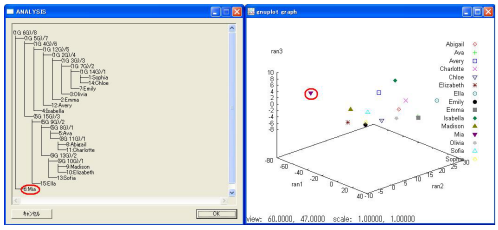

図 13: Left:dendrogram for students Right:3D plot by 3 subject score

1.5. 言語の類似状態 (トピック) を推定する LDA[20]

1.6. 時系列のパターンを推定する隠れマルコフモデル [21] 図 14 の左図は富士通の株価推移で，この隠れマルコフは 3 種の隠れモードを仮定している．右図の結果では一旦或るモードになると別のモードに移り難いことを示している．

図 14: Left:stock time series Right:Hidden Marcov mode

1.7. 異常値を検知する One Class SVM [22]

1.8. 推薦モデル GroupLense

図 15 の棒は 4 人の映画の評点を示す．左図では見ていない映画の評点は無い．右図は GroupLense が人と評点の類似性を見て評点を推定した結果である．見ていない映画の推定評点が高ければ推薦対象になる．

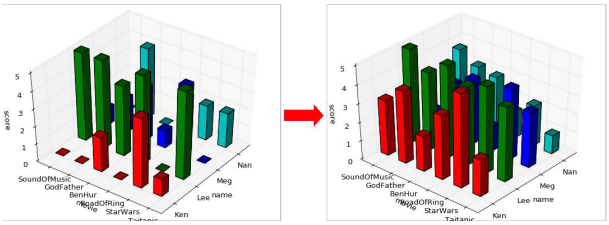

図 15: Left:exact score of movie Right:estimated score

2. グラフィカルモデル

提案ツールではデータから最適な関係をグラフで示すことができる．

2.1. ガウシアングラフィカルモデル

見かけ上の相関を排除した関係をグラフ表示する [23]　図 16 の左図の例は近代陸上 5 種競技データからの関係である．データ上は正の相関だが見かけ上の関係を排除すると負の相関が現れる．

2.2. 共分散構造 (SEM) モデル

項目間の関係を隠れた因子で説明するモデル [24] 図 16 の右図の例は社会活動のデータから社会的地位 (stage) を隠れ因子として説明した場合の関連の強さを表したグラフである．

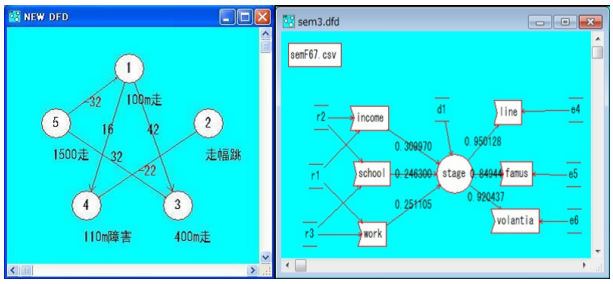


図 16: Left:5 athretic GGM Right: social status by SEM

### 2.3. ベイジアンネットワークモデル

データからベイジアンネットを自動生成する [26]

図 17 はローン債務者データからローン破綻 (def) を推定したベイジアンネットで．矢印に沿って確率伝播するので連結されたノードの条件に応じた確率が計算できる．

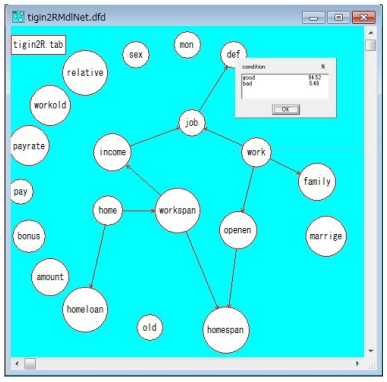


図 17: Baysian net for bank loan

### 2.4. 最短経路問題 (ダイクストラ法) [27]

図 18 の例は東京地下鉄の最短経路探索結果である．

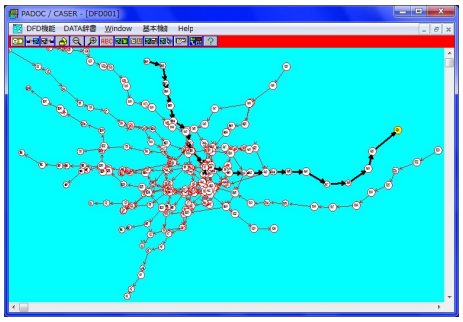

図 18: Optimize path in subway network of tokyo

### 2.5. 最大流量問題

図 19 の左図の例は各配管の容量制限内で配管網への最大流入量を示している．

### 2.6. 最大張木問題

図 19 の右図の例は全地点へ送電できる鉄塔の最短経路を示している．

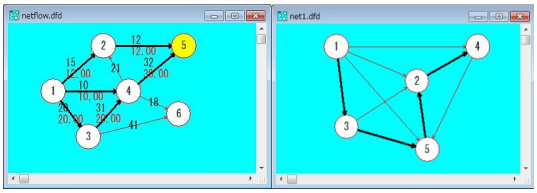


図 19: Left:max flow Right:max span tree

## 3. 最適化問題

提案ツールでは最適化問題として次の様な機能を提供している．

### 3.1. 線形計画問題

以下は実際の工場の人員配置問題を解いた例である．次の制約条件下で B 工場の最適な人員割当を求める．

＜ 制約条件 ＞

- A 工場は 8-17 時まで休まずに部品を作り順次 B 工場に搬送する
- B 工場は A 工場の部品の組立てと自工場でも部品製造と組立する
- 部品数以上に組立はできない
- 一日の生産数は，A 工場は 400 個 B 工場は 250 個
- 生産性は 1 人 1 時間当たり部品製造は 8.75 個 組立は 11 個
- B 工場では昼休みがあり 9 時台と 17 時台は 9 人 それ以外は 14 人出勤している

図 20 は B 工場の最適な人員割当の結果で，17 時台は 4 人省力化できることを示している．

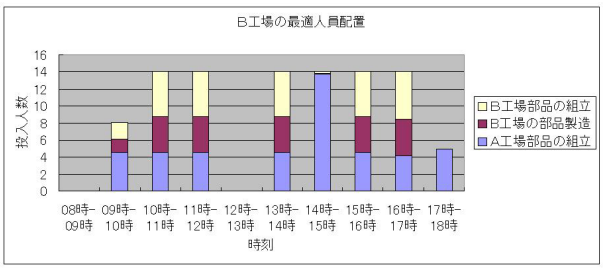


図 20: LienerPlan for personal allocation in factory

3.2. 整数計画問題 (分岐制限法) [28]

解が整数に限定される（ナップサック問題）

3.3. 非線形計画問題 [29]

図 21 の左図様な非線形問題を右図の様に条件式を設定すると最適問題を解くことができる．

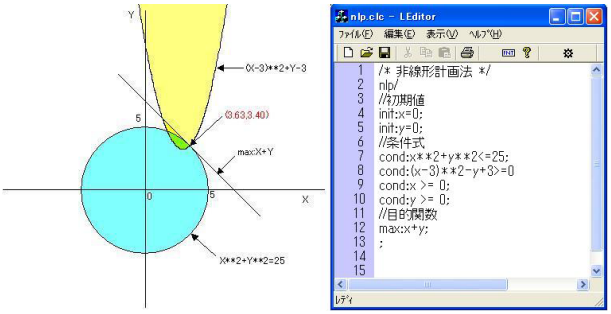


図 21: Left:promble Right:define nonleaner formula

3.4. 最適巡回路問題 (焼き鈍し法) [30]

図 22 は 3D のランダムな点を結ぶ最短順回路を解いた結果である．

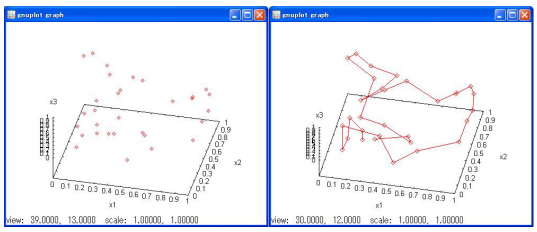


図 22: Left:random points Right:circuit path by anneal

4. 強化学習

最大報酬を得るために最適行動を学習するモデル [31]

図 23 の例は Sarsa 法による迷路探索の結果である [32]． この様に強化学習はゴールに達して初めて報酬が得られる場合でも解くことができる．

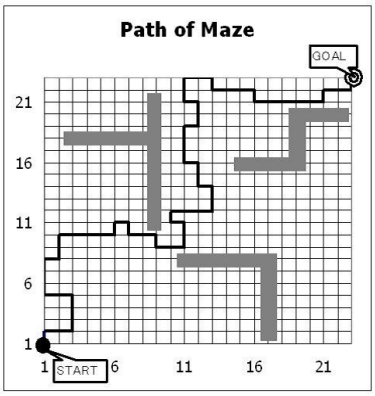


図 23: Reinforcement Learning for maze

## 8 考察

近年の機械学習やＡＩ関連の論文では殆どが Python で実装された実験報告になっている．これは Python がオブジェクト指向型の豊富な関数で複雑な数値計算ロジックを記述できるためと，直近の優良なモデルやライブラリィの公開が早く，これらを利用することで学術的な価値を高めるからである．逆に実用面としてデータの加工や編集の様な前処理をするには Python は記述が複雑すぎて一般的な技術者が利用するには負担が大きい．また現状では図 1 の右図で示す様なグラフィカル環境でモデルを構築する仕組みが Python には存在しない [3]．

実用的なデータ分析で要求されるのは高度な理論でなく，データ加工や編集を繰返して精度と頑健性があるモデルを構築し分り易く提示することである．提案ツール PADOC は平易な記述やグラフィカルな環境を提供し，試行の繰返しや評価を容易にしており実用的な面で優れていると考えられる．

## 9 まとめと今後

実用的なデータ分析として，全体像把握，仮説検定，比較検討，知識発見について，提案ツール PADOC を検討した．全体像把握については前処理大全 [2] の記述に沿ってデータ編集ツールとしての優位性を示し．比較検討，仮説検定，知識発見については提案ツールが提供するモデルがグラフィカルな環境によって性能をよく表す事を示した．

2017 年 DeepMind 社による棋譜学習しないでプロ級になる囲碁の学習モデル AlphaGoZero が発表された [33]．この論文の最後には「人類が数千年かけて習熟した囲碁の技を我々は数日で達成した」とある．機械学習の理論やモデルは革新的なものが日々研究され

ていると考えられる．この様なモデルをいち早く取り込める実用的なデータ分析環境として提供していきたいと考える．そのためPythonで作られたライブラリィを取り込むＡＰＩを公開して多くの人がこの環境をカスタマイズできる様にしたいと考えている．.

## 参考文献

[1] 孔子:『論語』為政第二 子曰．温故而知新．可以爲師矣

[2] 本橋智光, 前処理大全, 技術評論社 (2018)

[3] SAS Institute Japan 株式会社,SAS について https://www.sas.com/ja_jp/company-information.html

[4] IBM SPSS ソフトウェア https://www.ibm.com/analytics/jp/ja/technology/spss/

[5] 株式会社 NTT データ数理システム,S-PLUS for Windows https://www.msi.co.jp/splus/products/win/index.html

[6] MathWorks, 数学，グラフィックス，プログラミング https://jp.mathworks.com/products/matlab.html

[7] 木内貴弘,CDISC(Clinical Data Interchange Standards Consortium) 標準の概要 http://www.umin.ac.jp/indice/cdisc2009/01CDISC20091210pdf.pdf

[8] Tom White,Hadoop,Orailly(2013)

[9] データ分析競争サイト, https://www.kaggle.com/competitions

[10] 総務省,「高度 ICT 利活用人材育成プログラム開発事業（実践偏)」 データ解析手法とツール (2012) http://www.soumu.go.jp/main_sosiki/joho_tsusin/joho_jinzai/

[11] 坂本慶行 石黒真木夫 北川源四郎, 情報量統計学 § 6 分割表解析モデル, 共立出版 (1998)

[12] 長井章夫, マクロ経済指標を使った重回帰分析による倒産予測 SAS ユーザ会 2012

[13] 日本銀行, 時系列統計データ検索サイト, https://www.stat-search.boj.or.jp/

[14] John C. Platt,Fast Training of Support Vector Machines using Sequential Minimal Optimization

[15] 丹後俊朗, ロジステック回帰分析, 朝倉書店

[16] 中村剛,Cox 比例ハザードモデル, 朝倉書店

[17] Trevar Hastie,The Elements of Statical Learning,Springer(2008)

[18] 伊庭幸人, マルコフ連鎖モンテカルロ法とその周辺, 計算統計 2 , 岩波書店 (2005)

[19] Christopher M. Bishop:Pattern Recognition and Machine Learning § 11, Springer(2006)

[20] 持橋大地, 確率的トピックモデル http://www.ism.ac.jp/~daichi/lectures/H24-TopicModel/ISM-2012-TopicModels-daichi.pdf

[21] Mark Stamp:A Revealing Introduction to Hidden Markov Models(2012)

[22] 高畠泰斗 香田正人, 1 クラス SVM と近傍サポートによる領域判別 (2006)

[23] Hirono,Graphical Model http://www.mayomi.org/lecture01/BSJ2008/BSJ2008tutorial1.pdf

[24] 豊田秀樹, 共分散構造解析 [入門編], 朝倉書店

[25] 北川源四郎:時系列解析プログラミング—FORTRAN77 (岩波コンピュータサイエンス)

[26] 中井眞人, データからのベイジアンネットワーク構造推定, 人工知能研究会 2014

[27] 佐藤公男, グラフ理論入門, 日刊工業新聞社 (2004)

[28] 大山達雄, 最適化モデル分析, 日科技連出版社 (1993)

[29] 矢部博, 最適化とその応用, 数理工学社 (2006)

[30] Willam H. Press, Neumerical recipes Third edition § 10 (2011)

[31] Richard S. Sutton and Andrew G. Barto1,Reinforcement Learning:An Introduction

[32] 中井眞人,Sarsa(λ) 法による強化学習の汎用化 (2015) https://www.slideshare.net/MasatoNakai1/ros-63464994

[33] David Silver,Mastering the game of Go widthout human knowlege ,Nature(2017)

[3]但し深層学習ネットワーク構築用のグラフィカルツールとしてTensorBoard がある