# インタラクティブデータマイニングのための タッチデバイス用ユーザインタフェースの試作

加藤 大志[1]　本橋 洋介[1]　藤巻 遼平[2]　森永 聡[1]

**概要**：大規模なデータから有用な知識を発見するデータマイニングではアルゴリズムの処理に計算時間がかかる．これに対して，人の介入を前提として，人とアルゴリズムが連携しつつ処理を進めるインタラクティブデータマイニングという手法がある．アルゴリズムの処理中に人が介入するためには，それを支援するユーザインタフェースが重要となる．我々は，データマイニングにおける属性選択の問題に対して，FoBa アルゴリズムと連携してユーザが指示を与えるためのユーザインターフェースを試作した．データ分析者の試用により，操作に応じて瞬時にデータやモデルを可視化したグラフを更新することは試行錯誤に有効であることが分かった．

# A User Interface Prototype Using Touch Device for Interactive Data Mining

Daishi Kato[1]　Yosuke Motohashi[1]　Ryohei Fujimaki[2]　Satoshi Morinaga[1]

***Abstract:*** Data mining that finds useful knowledge out of large data requires lots of computation time to run algorithms. Interactive data mining is one style of data mining that involves human interference to running algorithms. To realize interactive data mining, a user interface is important to support such human interference. We focused on feature selection problem which is one of problems in data mining, and developed a prototype system to support interaction between a human and an algorithm called "FoBa." Our finding is that the interactive user interface helps a user to repeat trial-and-error data mining.

## 1. はじめに

データマイニングはデータから有用な知識を発見する手法である．この手法を用いると，データから規則性や少ないパラメータで表現できるモデルを，人手を介さずアルゴリズムで見つけ出すことができる．しかし，データマイニングには，データが大規模になると計算量が増え，結果を得るまでに時間がかかるという課題がある．そこで，データマイニングの研究では，計算量を抑えるアルゴリズムを考案するアプローチが多い．ところで，データマイニングの手法を用いてデータ分析を実行するデータ分析者は，納得のいくモデルを見付けるために試行錯誤を重ねる．例えば，センサーから得られた生データを加工してアルゴリズムを実行し，結果を見て再度別の加工をするといったことを行う．しかし，アルゴリズムの実行に時間がかかりすぎると，気軽に試行を繰り返すことができないという問題がある．これに対して，データマイニングにおいて人手を介すことを前提にして，人とアルゴリズムが連携する手法としてインタラクティブデータマイニング [2], [3]．がある．

インタラクティブデータマイニングでは，人がアルゴリズムに介入する．すなわち，アルゴリズムが出力する情報をもとにアルゴリズムの実行が完了する前にデータ分析者が指示を与える．このとき，データ分析者はアルゴリズムの出力を即座に理解し，アルゴリズム上矛盾がなく適切な指示を与えなければならない．また，アルゴリズムの計算処理を止めずに実行途中の状態で指示を与えて方向修正したい場合もある．このように人がアルゴリズムに介入することは容易ではなく，それを支援するためのユーザインタ

[1] 日本電気株式会社
NEC Corporation
[2] NEC Laboratories America, Inc.

フェースが望まれる．

我々は，データマイニングにおける問題の一つである属性選択問題において，FoBa アルゴリズム [1] と連携するユーザインタフェースを試作した．本稿では試作したユーザインタフェースの特徴と，実際にデータ分析者が試用した結果分かった効果について述べる．

## 2. 関連研究

KDD 2014 では Interactive Data Exploration and Analytics (IDEA) [*1] というワークショップが開催され，インタラクティブデータマイニングに関する発表があった．

文献 [4] には，近年盛んに研究されているビジュアルデータマイニングに注目してデータ可視化とデータ分析を組み合わせたツールが述べられている．開発された GNoT というツールは，SQL のようなクエリ言語を機能学習アルゴリズム用にも用意し，GUI でクエリを生成できるようにしている．GNoT は，人の指示に応じてインタラクティブに可視化するツールであり，我々が試作したツールと形態が類似している．

文献 [5] は，ユーザがモデルの中身を知らなくてもビジュアル表現を使ってインタラクティブに知識を入力するコンセプトを提案し，その実現に向けた研究課題を述べている．研究課題には，ユーザインタラクションのためのインタフェースと，それをもとにしたユーザモデルとデータモデルの推論，さらにそれらに適応した計算と可視化が含まれている．この中で我々の取り組みは，ユーザインタラクションのためのインタフェースと可視化に関係する．

一方，文献 [6] では，インタラクティブデータマイニングの潜在的な問題点を指摘している．すなわち，ユーザであるデータ分析者の意図が入りすぎることにより，得られる結果が予想を超えるものにならず，データマイニングの定義である「発見」に至らない可能性があると指摘している．本文献はいくつかの解決策が示唆しており，その中で我々は，特にユーザインタフェースの観点からユーザの制御可能範囲の適切な設定が重要であるという主張に同意する．

このようにインタラクティブデータマイニングの研究が進められているが，アルゴリズムと連携するためのユーザインタフェースに関する研究は少なく，本稿ではその一例として試作したユーザインタフェースについて述べる．

## 3. 試作システム

我々が試作したシステムは，属性数が極めて多い場合における重回帰分析の属性選択問題に関する．この問題を機械学習のアプローチで取り組むアルゴリズムの一つに FoBa [1] がある．試作システムでは，FoBa アルゴリズムによる推薦をユーザに提示し，ユーザからの入力により処理を進める．

図 1 に試作システムのユーザインタフェースのスクリーンショットを示す．画面は上下に 3 分割され，上から順に，操作の履歴に関する可視化エリア，モデル関する可視化エリア，操作エリアとなっている．ユーザは，中央および上の可視化エリアを見つつ下の操作エリアで操作をする．操作エリアでは属性がドラッグ&ドロップ可能なノードとして表現され，ユーザは左右のエリアに属性を移動することで属性の選択状態を指示する．アルゴリズムからの推薦はノードの色で表現され，より暖色のノードが選択すべきノードであることを示唆する．

本ユーザインタェースの開発において特に注意した点を列挙する．また，データ分析者からのフィードバックや効果的な使い方についても述べる．

- **特徴 1** ドラッグ&ドロップによる直感的な操作
  タッチデバイスを想定した場合は，ユーザからの操作はドラッグ&ドロップの方が容易であることが分かった．当初はクリックによる操作で実装したが，誤操作があったり，視覚的に何が起こるか分かりにくいなどの問題があった．
- **特徴 2** アルゴリズムからの推薦を視覚的に表示
  アルゴリズムからの推薦をユーザに押し付けがましくなく自然に提示する方法として，色による表示法を用いた．当初は推薦順に並べるという位置による表示法で実装したが，自動で位置が変わるとユーザが困惑するという問題があった．
- **特徴 3** 変化を瞬時に見せるための事前計算
  ユーザの試行錯誤を促進するには，ユーザの操作に応じた情報の変化が瞬時に見えるとよい．そこで，操作エリアでドラッグ開始すると，ドロップする以前にどのような変化が起こるか見せるようにした．これによりドロップ操作を止めたり，試しにドラッグを開始してみるなどの使い方が可能になった．これを実現するために，操作を 1 ステップ先読みする事前計算を行っている．
- **特徴 4** 事前計算途中のインタラクティブ操作
  上記の事前計算において 1 ステップ先の操作の候補が多い場合は，計算に時間がかかる．そこで，事前計算をユーザが操作する可能性が高い順に行うとともに，事前計算が終了していなくてもユーザの操作ができるようにした．これにより，ユーザが自分の意向で操作したい場合はすぐに操作し，システムからより豊富な情報を取得したい場合はしばらく待つという使い方が可能になった．

上記以外にも現バージョンに至るまでに得られた知見として 2 点補足する．一つは，データ分析者は可能な限りアルゴリズムによる最適化を期待することである．当初のバージョンでは，データ分析者の制御範囲を広げるため，

[*1] http://poloclub.gatech.edu/idea2014/

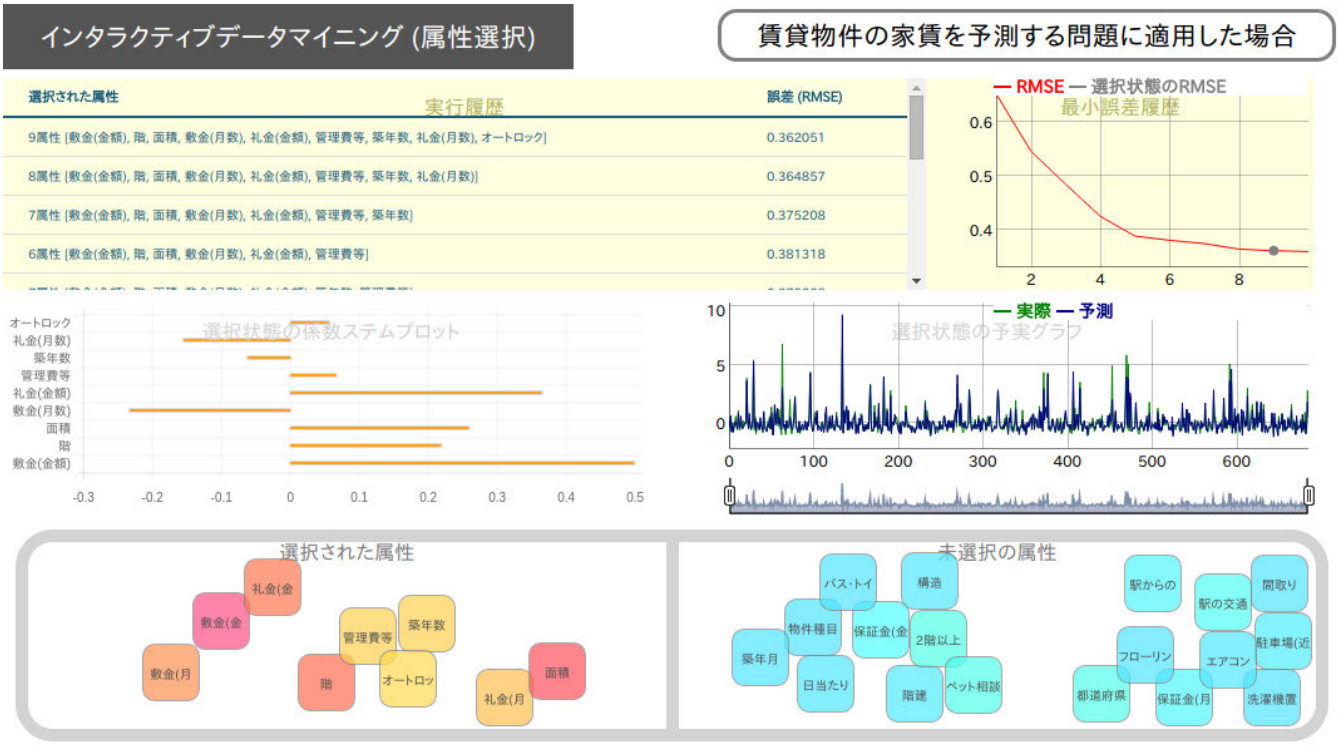


図 1　試作システムのスクリーンショット
**Fig. 1**　Screenshot of the prototype system

アルゴリズムからの推薦は控えめであった．しかし，データ分析者の要望はアルゴリズムによる最適化の結果をもとにして操作することであり，現バージョンでは，初期状態は時間かかってもアルゴリズムによる最適化を実行するように変更された．もう一つは，データ分析者はアルゴリズムによるブラックボックス的な推薦だけではなく，それを裏付ける様々な情報を統合的に見られることを要望することである．そこで，データに関する情報をできるだけ加工されていない状態で見る機能を追加した．例えば，図 1 の中央エリアの右に配置されたグラフ (以下，予実グラフ) は，入力データの実際の値とモデルによって予測された値が個別のデータ点としてプロットされている．この予実グラフも操作エリアのドラッグ操作に応じて瞬時に変化するため，属性の選択によってどのデータの予測値の誤差が小さいかもしくは大きいかなどが分かるという利点がある．

## 4.　試用例

本節では前節で説明した試作システムを実際にデータ分析者が試用した場合の操作の流れを説明する．試用に参加したデータ分析者は，データ分析業務経験 2 年の 1 名である．使用したデータは賃貸物件の家賃に関するデータで，予測対象の家賃以外に 27 の属性を持つデータである．

(1) 初期状態 (アルゴリズムによって最適化された状態) を見た
(2) 家賃に直接関係する属性は予測に用いたくないと考え，操作エリアで 3 つの属性を左から右にドラッグ&ドロップした
(3) 属性を追加することで平均誤差を下げたいと考え，操作エリアの右のエリアの色が濃い属性を 3 つ左にドラッグ&ドロップした
(4) 下記を確認しつつ，属性の入れ替えを試行錯誤した
- 属性の係数を確認するため，係数の正負とその大きさを視覚化したステムプロットを見た
- 予測の誤差を細かく確認するため，予実グラフをスライダーで拡大し，スクロールして見た
- 異常値により誤差が大きくなっているかを確認するため，予測値が外れているところを予実グラフを拡大して，探した
- 過去に選択した属性であるか確認するため，実行履歴の行を選択した
- 詳細な平均誤差の変化を確認するため，最小誤差履歴のグラフを拡大した

(5) 平均誤差が安定しある程度より下がらなくなったところで終了した

試用を通して分かった試作システムの特徴的な効果について述べる．(2) では，アルゴリズムが選択した属性をデータ分析者が不要と判断して外した．本データでは属性数が少なく事前に不要なデータを列挙することも可能であったが，属性数が非常に多い場合は後から外すという操作は有効であろう．(4) の属性の入れ替えの操作では，ドラッグの

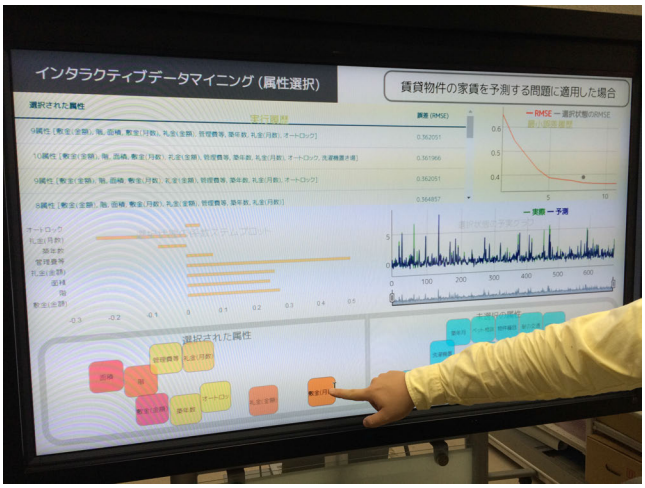


**図 2**　タッチデバイス付き PDP での操作シーン
**Fig. 2**　Usage scene of PDP with touch device

開始時点でグラフを確認し反対のエリアのドロップまで行わないケースがあり，事前計算による変化の可視化が試行錯誤に有効であることが分かった．また，グラフを拡大して詳細を確認するなど，生データをすぐに確認できるインタラクティブなグラフも効果があった．また，履歴をすべて残すことで過去の操作を振り返ることも効果的であった．

今回の試作システムはタッチデバイスで操作することを想定した．図 2 に試作システムをタッチデバイス付き PDP で用いたときの操作シーンを示す．タッチによるノードの操作は直感的であり，試行錯誤の作業の苦労が軽減されることが期待できる．

今回の試作システムでは実装されなかったが有効でありそうな機能としては，ユーザの操作をもとに再度アルゴリズムによる最適化を実行する機能がある．すなわち，アルゴリズムの 1 ステップ毎にユーザが介入するのではなく，ある程度収束するまでアルゴリズムを走らせ，その結果をユーザが操作するという反復を繰り返す方法が考えられる．

## 5. おわりに

本稿では，インタラクティブデータマイニングにおいてユーザとアルゴリズムが連携するユーザインタフェースの課題に着目し，属性選択の問題に関して試作したシステムについて述べた．試作システムをデータ分析者に試用してもらうことで，いくつかの点で有効性を確認した．有効性のさらなる確認には，スキルの異なるデータ分析者による評価やより大規模なデータでの評価が必要となるだろう．また今後の発展としては，属性選択の問題以外での適用や複数の問題に共通したユーザインタフェースの開発があるだろう．ユーザインタフェース以外にも，ユーザの操作をもとによりよいモデルを作るアルゴリズムの開発が望まれる．これらの研究が進めば，インタラクティブデータマイニングはより実用的なものになるだろう．

## 参考文献

[1] Zhang, T.: *Adaptive Forward-Backward Greedy Algorithm for Sparse Learning with Linear Models*, Advances in Neural Information Processing Systems 21 (2009).
[2] Ankerst, M.: *Human Involvement and Interactivity of the Next Generation's Data Mining Tools*, Proc. Workshop on Research Issues in Data Mining and Knowledge Discovery (2001).
[3] Zhao, Y., Chen, Y. and Yao, Y.: *User-centered Interactive Data Mining*, Proc. of the IEEE-ICCI'06 (2006).
[4] Ganeshapillai, G., Brooks, J. and Guttag J.: *Rapid Data Exploration and Visual Data Mining on Relational Data* Proc. of the KDD 2014 workshop on Interactive Data Exploration and Analytics (2014).
[5] Endert, A., North, C., Chang, R. and Zhou, M.: *Toward Usable Interactive Analytics: Coupling Cognition and Computation* Proc. of the KDD 2014 workshop on Interactive Data Exploration and Analytics (2014).
[6] Miettinen, P.: *Interactive Data Mining Considered Harmful (If Done Wrong)* Proc. of the KDD 2014 workshop on Interactive Data Exploration and Analytics (2014).