6L-07

大規模データ分析のための可視化手法に関する検討

野田昌太郎 † 河井悠佑 †† 趙セイ ††† 杉浦健人 ††† 石川佳治 ††

† 名古屋大学工学部電気電子・情報工学科 †† 名古屋大学大学院情報学研究科

††† 名古屋大学大学院情報科学研究科

1 はじめに

近年，コンピュータで扱われるデータサイズが爆発的に増加しており，大規模データ分析の需要が高まっている．インターネット上のサイト，スマートフォンのアプリケーションなどを筆頭に，大規模データを得られる機会が多くなり，それらを分析して見識を得るための分析手法に関する研究も盛んに行われている．

大規模データ分析への需要が高まったことで，データベース上で分析を行う in-database アナリティクス [1] が注目を集めている．In-database アナリティクスはデータベースからデータを取り出さずに分析を行う手法である．この手法では，分析処理をデータベースのライブラリとして実装するため，他のソフトウェアにデータを渡す必要がなく，高速な分析処理が可能になる．

また，大規模データ分析のための可視化フレームワークも重要性を増している． サイズが大きいデータの分析は，情報量が多すぎてユーザが直感的に理解できる代物ではなくなってしまっているためである．例えば，文献 [2] では，可視化インターフェースによる機械学習モデルの比較が提案されており，要素ごとのドリルダウン，ロールアップなどの操作がサポートされている．

本研究では in-database アナリティクスを用いた大規模データ分析の可視化インターフェースを提案する．In-database アナリティクスは大規模データを分析する優れた処理基盤であるが，そのインターフェースはコマンドラインのみであり，使用には専門的な知識が必要になる．そこで可視化インターフェースをフロントエンド，in-database アナリティクスをバックエンドに持つフレームワークを提案し，大規模データの直感的な分析を可能とするシステムの実装を目指す．

2 In-database アナリティクス

In-database アナリティクスは RDBMS からデータを取り出さずにデータ分析を行う技術のことで，RDBMS に分析処理のライブラリを組み込むことで，データベースのプロセス空間内での分析処理を可能とする．また，データベース上で実行するため，選択や射影などのデータベースらしい機能を取り入れつつ分析可能なことも大きな利点である．

本研究では，in-database アナリティクス用ライブラリとして MADlib [3] を使用する．MADLib は Greenplum や PostgreSQL などの RDBMS 向けにに開発された，統計解析や機械学習のアルゴリズムを集めたライブラリである．ライブラリはすべて SQL の関数として実装されており，分類，回帰，クラスタリングなどのアルゴリズムが用意されている． 例えば，MADlib で心不全の患者データ “patients” テーブルを分析する場合について考える．このデータは患者が 1 年以内に心臓発作を再発したかどうかを示す “second_attack”，患者が anger control の治療を受けたかどうかを示す “treatment”，特性不安の係数 “trait_anxiety” の 4 属性を持っている．ここでは “treatment”，“trait_anxiety” 要素から “second_attack” が発生するかどうかをロジスティック分析で予測する．まずはデータを学習するために，以下の SQL を実行する．

```
SELECT    madlib.logregr_train(
  'patients',
  'patients_logregr',
  'second_attack',
  'ARRAY[1，treatment，trait_anxiety]'
);
```

“patients” は学習するテーブル，“patients_logregr” は学習モデルを出力するテーブル，“second_attack” は学習モデルで予測するラベル，“ARRAY[...]” は学習に用いる特徴量である．学習が完了すると，“patient_logregr” テーブルに学習モデルが作成される．モデルを用いた予測は，以下の SQL で行う．

```
SELECT madlib.logregr_predict(
  coef,
  ARRAY[1，treatment，trait_anxiety]
)
FROM patient_test p，patient_logregr m;
```

“coef” は作成したモデル内の属性で学習に使用した各属性の係数，“ARRAY[...]” はモデル作成に用いた配列，

A Visualization Method for Large-scale Data Analytics

Shotaro Noda†, Yusuke Kawai††, Zhao Jing†††, Kento Sugiura†††, and Yoshiharu Ishikawa††

†Department of Information Engineering, School of Engineering, Nagoya University

††Graduate School of Informatics, Nagoya University

†††Graduate School of Information Science, Nagoya University

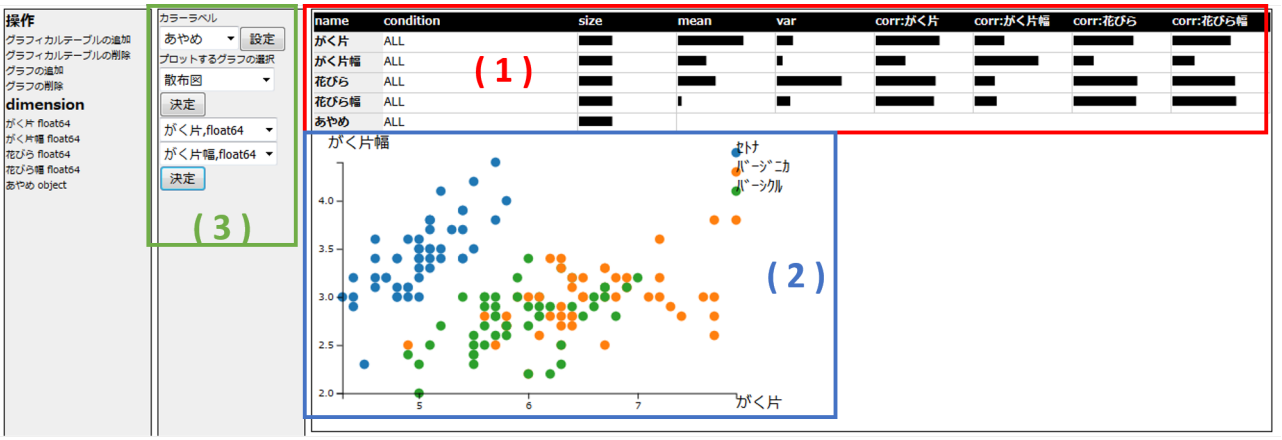

図 2　可視化インターフェースの画面

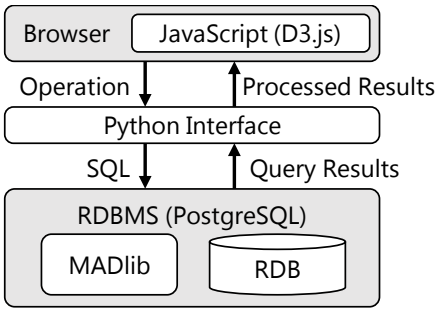

図 1　提案フレームワーク図

“patients_test” は予測に用いるデータである．このように SQL から関数として機械学習処理が呼び出せるため，射影や選択などの SQL のその他の処理と組み合わせることが可能である．

3　可視化手法の実装

提案フレームワークは図 1 に示すようになっており，ユーザがブラウザ上で行った操作に応じてサーバ側で PostgresSQL 及び MADlib による分析が行われ，分析結果をブラウザ上に可視化する．ブラウザ上での可視化には JavaScript 言語のライブラリである D3.js を用いており，単純に分析結果のテーブルを表示するだけでなく，集計値のグラフィカルテーブルとしての表示や散布図の描画など，分析結果の理解を補助するよう可視化する．なお，サーバ側のインタフェースは Python 言語により実装しており，ブラウザ上で行われた操作に応じた分析用 SQL の生成，必要に応じた問合せ結果の加工などを担う．

可視化インターフェースの例として，Edgar Anderson のあやめのデータセット [4] を可視化したものを図 2 に示す．このデータは 3 種類のあやめ (セトナ，バージニカ，バーシクル) のがくの長さと幅，花びらの長さと幅のデータである．現在のインターフェースでは，指定したデータに対してグラフィカルテーブルと散布図の描画を行う．グラフィカルテーブルは図 2 上部 (1) に示されており，各属性の平均，分散および属性間の相関を棒グラフで表示する．散布図は図 2 下部 (2) に表示されており，散布図のラベルや軸の設定は図 2 左側のメニュー (3) で行う．例えば図 2 では軸としてがく片とがく片幅を，カラーラベルとしてあやめの種類を使用している．

4　おわりに

本研究では，in-database アナリティクスを用いたデータ分析の可視化により，大規模データを対話的に分析できるフレームワークを検討した．今後はより直感的なインターフェースを実装し，機械学習の解析，統計解析をよりスムーズに行えるようにすることが課題である．

謝辞

本研究の一部は，科研費（16H01722）および CREST「大規模・高分解能数値シミュレーションの連携とデータ同化による革新的地震・津波減災ビッグデータ解析基盤の創出」による．

参考文献

[1] X. Feng, A. Kumar, B. Recht, and C. Ré, “Towards a unified architecture for in-rdbms analytics,” in *SIGMOD*, pp. 325–336, ACM, 2012.

[2] M. Kahng, D. Fang, and D. H. P. Chau, “Visual exploration of machine learning results using data cube analysis,” in *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, HILDA ’16, pp. 1:1–1:6, ACM, 2016.

[3] MADlib : `http://madlib.apache.org/` (accessed: January 5, 2018).

[4] UCI Machine Learning Repository Iris Data Set : `https://archive.ics.uci.edu/ml/datasets/iris` (accessed: January 5, 2018).