Ｗｅｂデータ対応リアルタイムデータマイニングツールの開発
―データマイニングを身近なものに―

１．　背景

大量のデータから、隠れた知識を生成する方法として、最近データマイニングという手法が注目されているが、一般にはあまり普及していない。

普及を阻んでいる理由としては、費用的な面と、ツールを扱うには統計の専門家やプログラミング力が必要とされることなどが考えられる。

また、Web 上の膨大なデータの中には、今まで気が付かなかった知識が隠れている可能性が高いが、Web から<Table>タグ内の表データを含め多様な形式のデータを取り込めるデータマイニングツールは見当たらない。

２．目的・ねらい

統計の専門化だけにとどまっているデータマイニングの市場を、一般層まで広げたい。また、従来のデータマイニングツールには無い、Web 上のデータを取り込める機能や、マクロやプログラミングを使わなくても Web アプリと連携して予測モデルを自動更新できる機能などを実現し、データマイニングの応用範囲を広げたい。

本ソフトはSAS社やSPSS社などから出されているような統合的なツールを目指すのでは無く、一般の人が気軽に使えて、データマイニングのメリットを享受できるものを目指す。しかも、実用に耐えるデータ量（エクセルレベル以上）を扱えて、効率的な分析アルゴリズムを備えるソフトにしたい。データマイニングの手法としては、当面は利用頻度の高い予想系の重回帰分析を採用し、将来はバスケット分析など他の分析機能を追加する。

３．　開発の内容

（１）HTML,CSV,TSV など様々なファイル形式で存在する表データを、インターネットやディスクなどから容易に入力し、分析結果を得ることができる。

分析対象データは簡単な操作で、格納場所（Web、PC）やフォーマットなどをあまり意識せずに取り込むことができる。取り込み後の分析から予想モデルの作成、分析結果の簡単な解説文出力までを本ソフトが一気に処理する。そのため、ユーザはマウスで予想対象列を指定するだけで、膨大なデータから傾向や法則などを容易に得ることができる。

なお、データ入力方法は次のとおりである。

●ファイルを画面上にドラッグ＆ドロップ（D&D）
CSV ファイルと TSV ファイル（タブ区切り）は拡張子によらず自動認識で読み込める。特殊な区切り文字にも対応している。
●インターネットショートカット（拡張子：'．URL'）を D&D
ブラウザのインターネットショートカットから本画面上に D&D すると、本ソフトの通信機能が自動的に働いてネット上のデータを収集できる。
●「貼り付け」機能を利用
他のソフトからコピー＆ペーストで表データをダイレクトに貼り付けることができる。
●プロジェクトファイルを利用
本ソフト専用のプロジェクトファイル（拡張子：'．DPR'）に、ネットから収集すべきファイル名（複数可）や FTP パスワード等を事前に記述しておくことにより、バッチ的にデータを取り込める。HTTP と FTP の両方に対応している。
●既存の表収集ソフト「＜Table＞バインダ」との連携
開発者作成の他のソフト「＜Table＞バインダ」との連携機能があり、これを利用して、HTML ファイル内の<Table>タグ内の表データを分析データとして取り込める。

（２）マクロやプログラミングを使わずに、容易に予測モデルの作成とそれを使ったシステムを構築できる。また、予測モデルの更新作業を自動化できる。

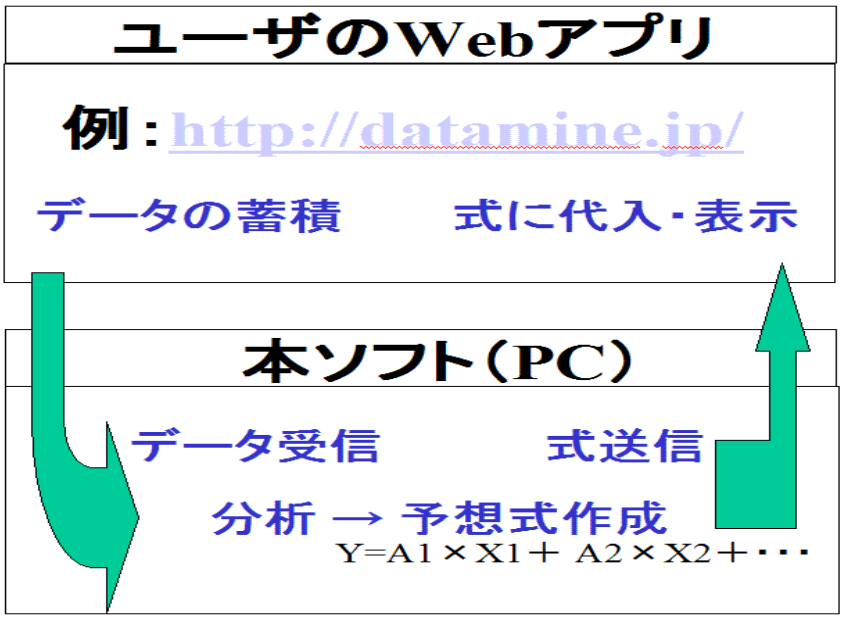

図１

Web アプリとの連携例として今回作成したのが、「当たるも八卦」（Web 側 TOP 画面 http://datamine.jp/ ）というシステムである。これは Web アプリ側（データ蓄積）と、リモート PC 側（予想モデル作成）の連携動作で動いている。（図１を参照）

Web 上で蓄積されたデータを、本ソフトに取り込み、分析して予想式（つまり偏回帰係数と定数項）を決定し、その係数ファイルを Web サーバーに送信する。この取り込みから送信までの一連の流れは、プロジェクトファイルで簡単に設定できる。ちなみに「当たるも八卦」は、OS のタスクスケジューラに設定されており、２日毎に予想式が自動更新されている。

この Web 連携機能を利用すると、Web 対応の分析予想系のシステムを構築する場合に、ユーザは Web 側のアプリを作成するだけで良い。但し、前処理がほとんど不要なデータを吐き出すように Web アプリを作成しておく必要がある。

（３）予測因子をＡＩＣ法等により適切にかつ自動的に選択し、予測信頼性の高い最適なモデルを作成できる。

説明変数同士の相関が強いデータが存在する場合、多重共線性の関係でそのまま分析すると予測信頼性が落ちるが、本ソフトでは重回帰分析処理の前に相関の強いデータの一方を自動的に外し、さらにＡＩＣ法と変量増加法の組み合わせで多重共線性を回避している。

（４）解析結果に悪影響を及ぼす異常データを除去する

数字以外のデータが入っている行を自動的に除外するようにした。

４．　従来の技術（または機能）との相違

（１）多様な形式の Web データを取り込める。

Web から<Table>タグ内の表データを含め多様な形式のデータを簡単な操作で収集できる。

（２）Web アプリとの連携機能がある。

本ソフトを分析、通信部品として利用すると、Web 側のアプリ開発とプロジェクトファイルの作成だけで分析・予想系の Web システムを容易に構築できる。

（３）操作が簡単である。

データ入力から分析結果の出力までの操作は、①Ｄ＆Ｄ、②[編集]ボタンクリック、③予想対象列クリック、及び④[個別分析]ボタンクリック

だけである。
※③の表画面では、必要に応じて表やデータを修正できる。
（４）分析結果を文章で出力する。
重回帰分析の結果を簡単な文章で出力する。
例：http://datamine.jp/report2-1.htm
（５）分析計算が高速である。
最適なモデルを自動選択するための独自のロジックを採用し、今回これをさらに改善した。図２において、縦軸は所要時間、横軸はデータ量である。左図は従来の手法、右図は独自の手法を使ってそれぞれ処理時間を計測したものである。

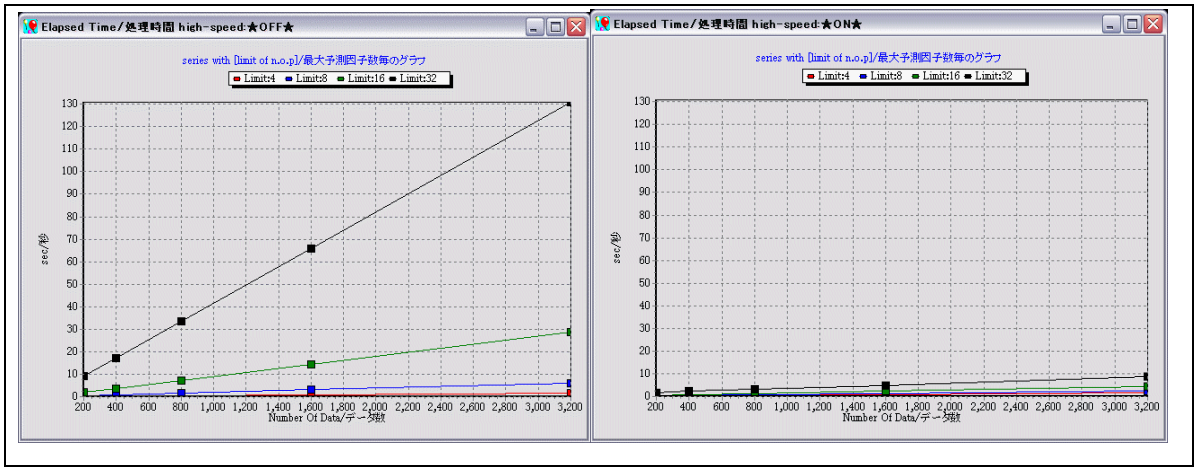

図２

５．　期待される効果

統計の専門化だけにとどまっているデータマイニングの市場が、一般層まで広がることが期待できる。また、Web アプリとの連携システムなど、データマイニングの応用範囲が今よりも広がると思われる。

６．　普及（または活用）の見通し

Web 上にシェアウェアとして公開し、だれでも気軽に本ソフトを利用できるようにしたい。本ソフトのダウンロード数は、開発者が作成した既存ソフトのダウンロード数の５万程度を見込む。

７．　開発者名（所属）

若松 桜男（有限会社ソフトポート）

（参考）開発者ＵＲＬ http://www.softport.co.jp/