

汎用データマイニングツール 「Visual Mining Studio」新機能紹介

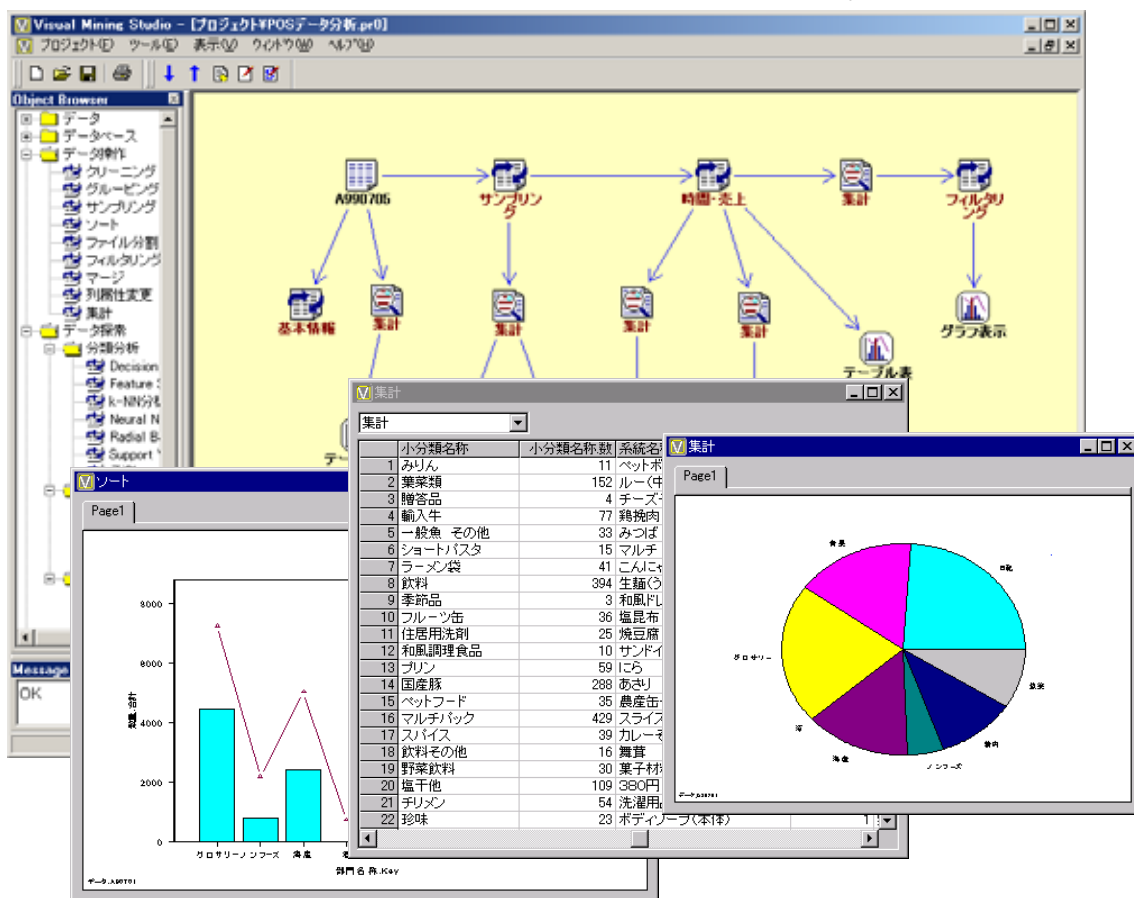
株式会社数理システム

データマイニング室 雪島正敏

vmstudio-info@msi.co.jp

1. 概要

Visual Mining Studio (以下 VMStudio) は株式会社数理システムによって独自に開発されたデータマイニングツールであり、データの前処理・加工から探索までの一連のデータマイニング作業をビジュアル的に行う事をサポートするツールである。



VMStudio はデータの加工等の前処理から探索までの一連のデータマイニング作業を視覚的にサポートするツールである。データ加工機能として、良く使われるサンプリング・クリーニング・ソート・マージ・集計などの機能から、データの細かな整理・加工を行うためのスクリプト機能まで提供する。データ探索機能として、代表的な手法のみならず、最新のデータマイニング手法を取り入れるとともに、ユーザの使いやすさにも工夫をしている。また、汎用データ解析ソフト S-PLUS と連動し、S-PLUS のもつ豊富な統計関数やグラフ表示機能を使用する事を可能にしている。本報告では、VMStudio の最新版(Ver.3.0)

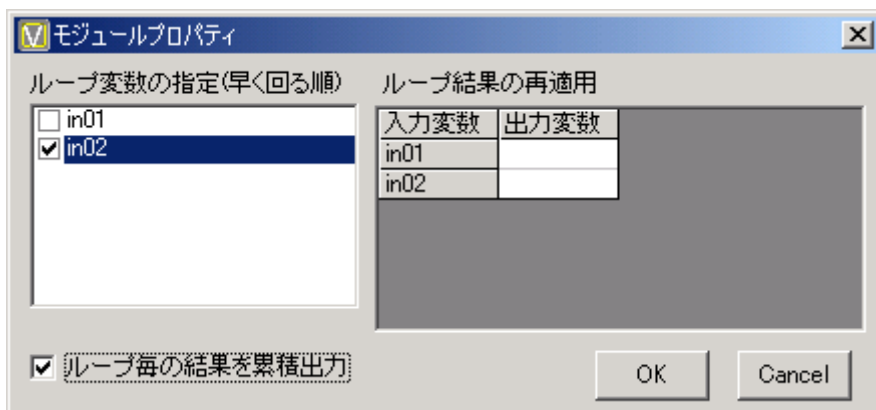
に含まれる機能を紹介する。

2. ビジュアルプログラミング機能

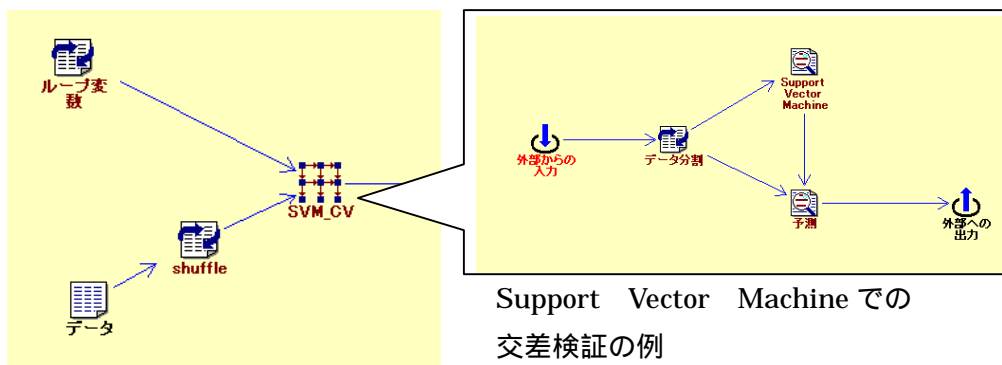
VMStudio は、ビジュアルプログラミングの有力な道具でもある。ビジュアルプログラミングは、図に示したように、データ分析に現れる基本要素(データ・分析手法)をアイコンで表現し、それらをプロジェクトボードに配置し、線で繋ぎ処理の流れを定義(プログラミング)する。各アイコンの処理の結果は、保存・非保存を選択する事が可能なので、処理の分岐等も効率的に行える。分析内容は、複数の処理フローからなるプロジェクトで管理する。同時に複数のプロジェクトを編集することも可能で、プロジェクト間でアイコンや処理フローのコピーも簡単に行える等、プロジェクト管理を容易に行う事が出来る。また、作成したプロジェクトをモジュール化し、他のプロジェクトで使用する事も出来る。新バージョンでは以下の機能が追加された。

2.1 ループ処理機能

ループ処理機能とは、モジュール化されたプロジェクトに必要なデータとループ変数を渡すことで、そのモジュールを複数回、実行する機能である。ループ処理では同一のデータに対して複数回の処理を行うことも、モジュールの出力を次のループ処理の入力にすることも可能である。また、スクリプトを用いループを途中で終了して抜けることも可能である。

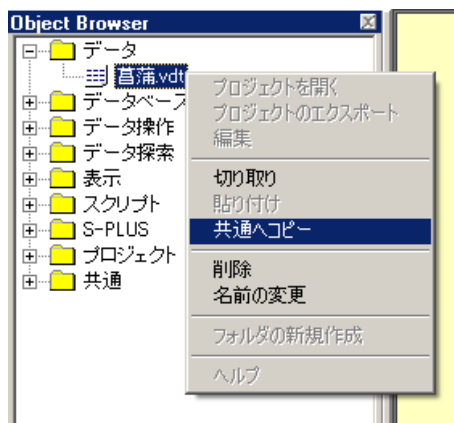


ループ処理を用いると、下図のように、交差検証を行う処理を簡単に実装できる。



2.2 ユーザ管理

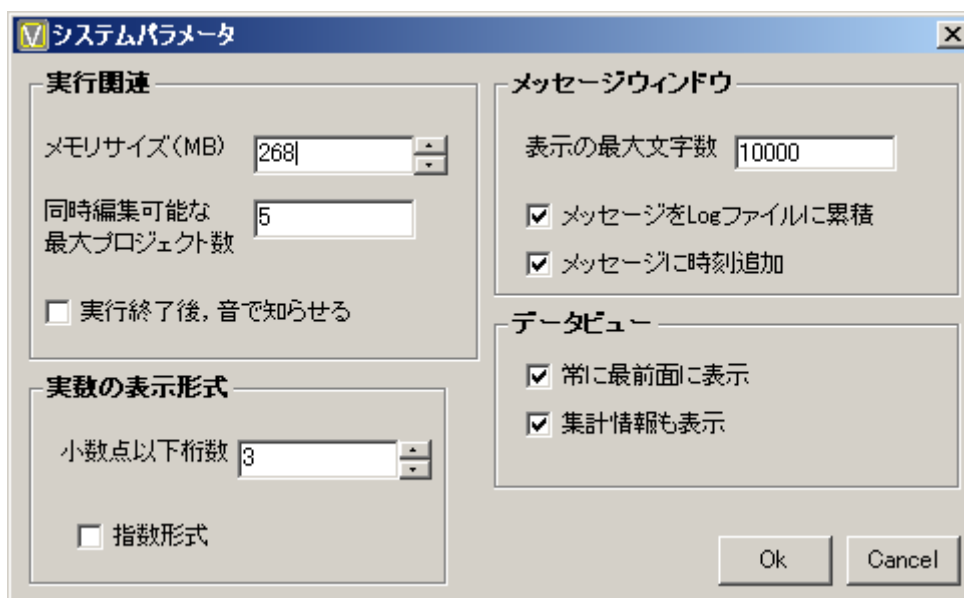
ユーザ毎にデータやプロジェクト等の管理が行えるようになった。これに伴いすべてのユーザがデータを置くことのできる共通フォルダの機能が新設された。これにより、ユーザごとに作業領域を分離しながら、ユーザ間でデータを交換したり、プロジェクトを共有したりして作業の効率を高めることができるようになった。



2.3 その他の操作性の向上

その他の操作性の向上として、下記のシステムオプションが追加された。

- メッセージウインドのログファイルへの出力機能
- 実行終了時に音で知らせる機能
- メッセージを出力した時刻を追加する機能
- データビューでの表示形式を指定する機能



3. データ加工機能

データのマイニング作業の70%~80%はデータの整理・加工処理に費やされる。そこで、VMStudioではデータの整理・加工処理の細かな制御を高速に行えることにも重点を置いている。新バージョンでは以下の機能が追加・改良された。

3.1 グルーピング機能

グルーピング機能は、簡単なマウス操作で、数値データをカテゴリ化したり、カテゴリデータを再カテゴリ化する機能である。



3.2 カテゴリデータの数量化機能

列属性の変更機能にカテゴリデータを数量化する機能を追加した。数量化では、カテゴリデータを 0-1 の複数の列に変換する。下図のように、文字列属性の列に関して数量化にチェックをすることで数量化した列を追加する。

列形式変更

属性指定

列名	属性	新列名	新属性	数量化
pick	文字列	pick	文字列	X
income	文字列	income	文字列	
moves	文字列	moves	文字列	
age	文字列	age	文字列	
education	文字列	education	文字列	
employment	文字列	employment	文字列	
usage	整数	usage	整数	
nonpub	文字列	nonpub	文字列	
reach.out	文字列	reach.out	文字列	
card	文字列	card	文字列	

HELP

OK Cancel

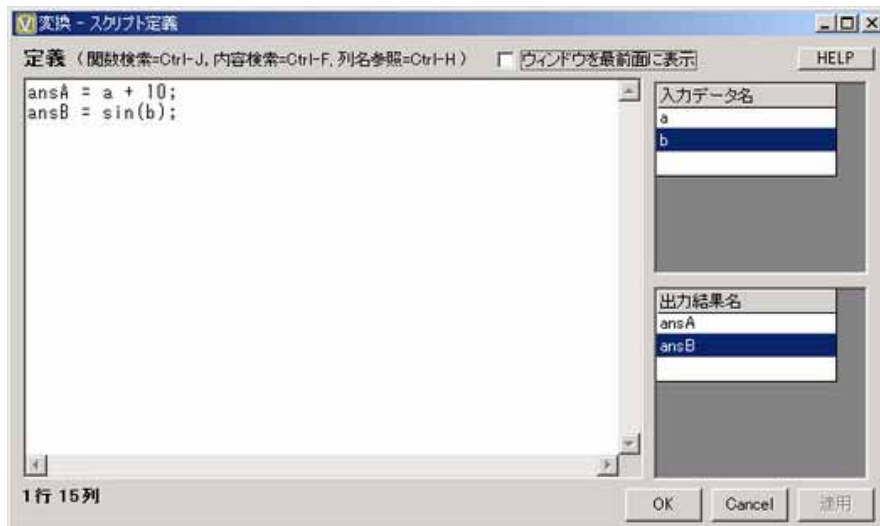
3.3 ソート機能 改良

ソートを行う時に、ソートキー列毎に昇順・降順の指定できるように改良した。



3.4 スクリプト機能 改良

入力データから列名を自動的に取得してスクリプトに挿入したり、また関数名一覧から関数を選択して、スクリプトに使用できるようになる等、スクリプトの編集機能を改善した。



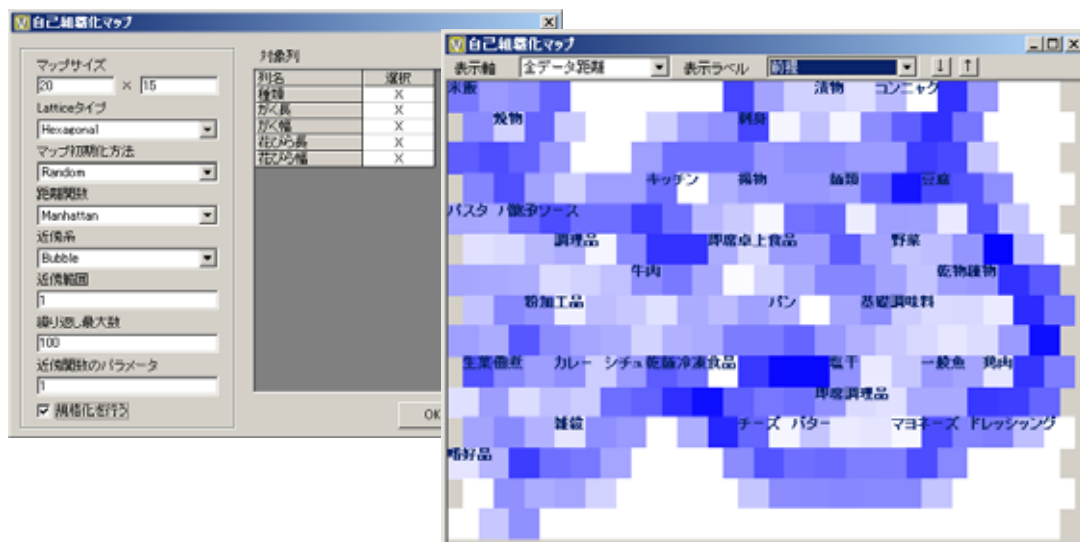
4. データ探索機能

膨大な量のデータの中から効率よく、有益な知識を得る事がデータマイニングの目的である。VMStudio では、出来るだけ広範にわたり、多くのデータマイニングの手法を提供するだけでなく、ユーザの使いやすさにも工夫している。新バージョンでは以下の新しい機能が追加された。

4.1 クラスタ分析機能

4.1.1 自己組織化マップ

自己組織化マップを作成する機能を追加した。この自己組織化マップでは、学習方法はバッチ学習と呼ばれる学習方法を用い、マップの初期化方法ではランダムと主成分分析を選択できる。また、マップの形態はRectangular と Hexagonal を選択できる。



4.1.2 クラスタリング機能 改良

すべてのクラスタ分析機能(K-Means,BIRCH,OPTICS,自己組織化マップ)に、下記の機能を追加した。

カテゴリデータの取り扱い

カテゴリデータを数量化して取り扱えるようにした。

入力データの規格化

入力データを規格化するかどうかのオプションを選択できるようにした。

4.2 分類分析機能

4.2.1 k-NN 分析機能 改良

k-NN 分析で回帰分析を行えるようにした。

4.2.2 Support Vector Machine 機能 改良

Support Vector Machine で多群判別を行えるようにした。

4.2.3 Decision Tree 機能 改良

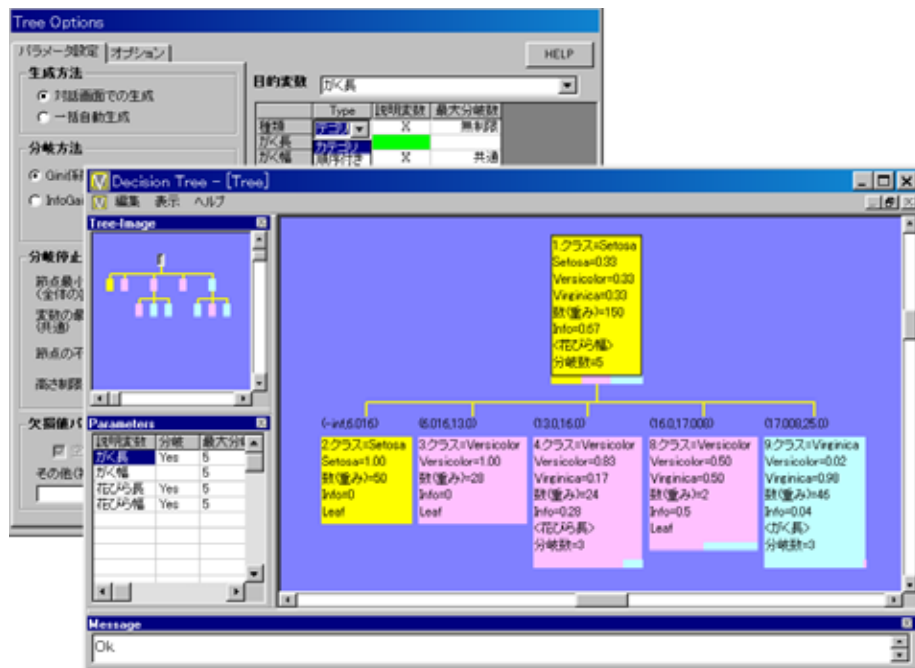
Decision Tree 機能に以下の機能を追加した。

順序付変数

カテゴリデータを順序付変数として取り扱う機能を追加

GUI 機能改善

GUI の機能を改善した。



4.2.4 重み付学習機能

全ての分類分析機能 (Neural Network, Decision Tree, k-NN 分析, Radial Basis Function Network, Support Vector Machine) に、重み付の学習機能を追加した。重み付学習機能とは、個々のデータに対する重要さ (重み) を指定して学習する機能である。重みの付け方は以下の通りである。

クラス毎に重みを付ける機能

クラス毎に重みを指定することで特定のクラスをより正しく学習する機能。例えば、希な (しかし重要な) クラスの重みを上げる事で、それらをより正しく識別するモデルを構築できる。

行毎に重みを付けて学習する機能

行毎に重みを指定することで、特定の行をより正しく学習する機能。例えば、昔のデータに比べ今のデータの重みを上げる事で、過去の情報を考慮しつつも、より最新の情報に重きをおいたモデルを構築できる。

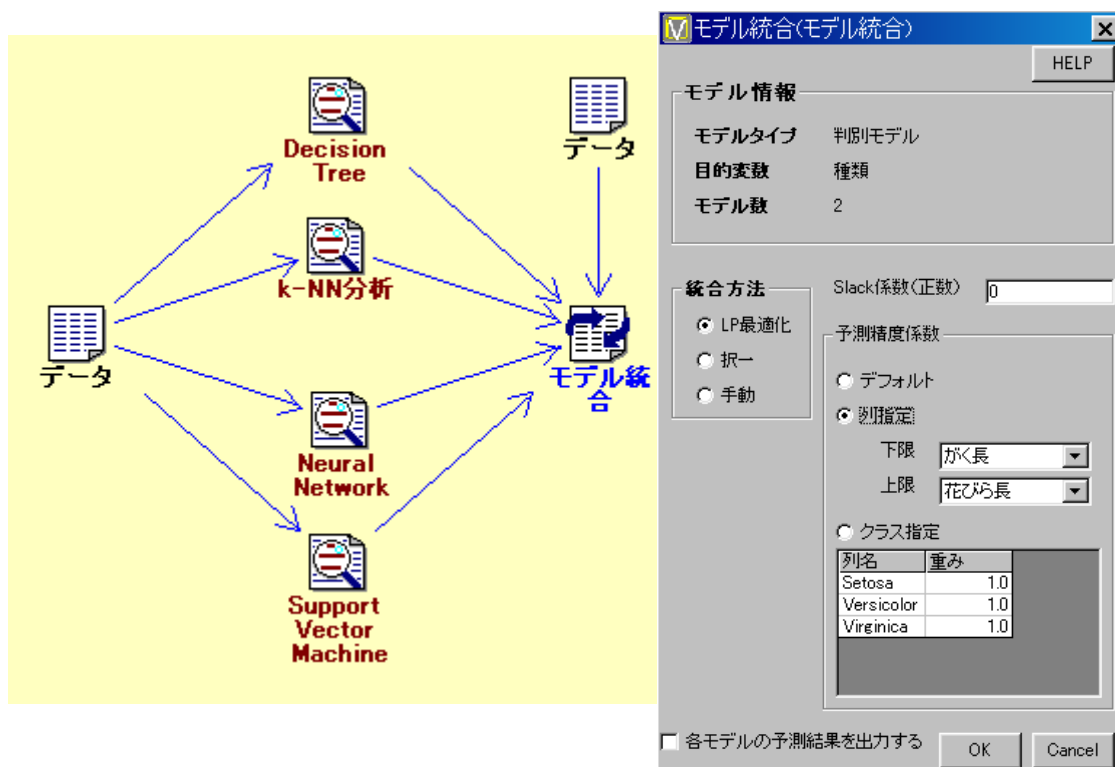
4.2.5 モデル保存機能

分類分析で作成したモデルを保存して、アイコンとして登録し、再利用する機能。

4.2.6 モデル統合機能

モデル統合機能は、Neural Network や Decision Tree 等、分類分析機能で作成した複数のモデルを統合し、より汎化性の高いモデルを作成する集団学習を行う機能である。この機能では、複数の学習アルゴリズムで作成したモデルから、

- 1) 線形計画法を用いてそれぞれのモデルに対する最適な重みを求める「LP 最適化」
 - 2) ひとつのモデルを選択する「択一」
 - 3) モデル毎に手動で重みをつけ、全体の結果を予測する「手動」
- を行うことが出来る。



4.3 アソシエーション分析機能

4.3.1 アソシエーション分析 改良

既存のアソシエーション分析を高速化した。

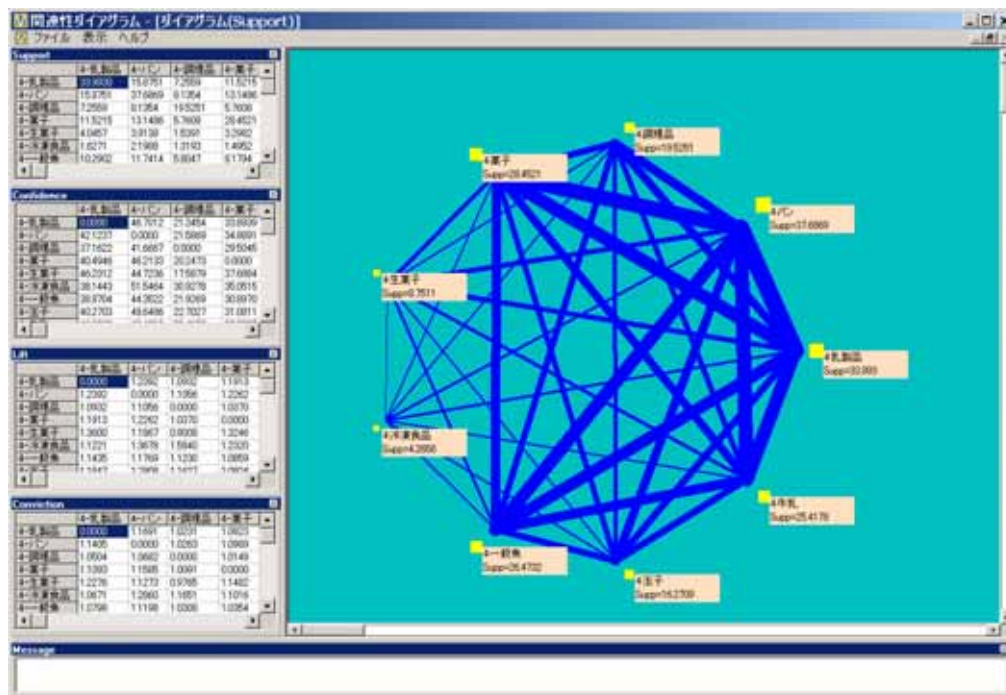
4.3.2 時系列アソシエーション分析

時系列データに対して、順序を考慮した事象間の関連性を求める時系列アソシエーション分析機能を追加した。

4.3.3 関連性ダイアグラム分析 改良

関連性ダイアグラム分析で下記の GUI 機能を改善した。

- 関連性ダイアグラム表示で、サポート、信頼度、Lift、Conviction それぞれの指標でグラフを描画する機能を追加した。
- ノード間の連結の強さに閾値を設けて表示する機能を追加した。
- ノード間の連結の詳細を表示する機能を追加した。



5. 表示機能

新バージョンでは以下の表示機能が強化された。

5.1 テーブル表示機能 改良

大規模データ（例えば、100 万件以上のデータでも）を分割することなく、一括「テーブル表示」を違和感なく行う機能を追加。また、メニューから「データビュー」を選択する事でデータ表示を行うことを可能にした。

6. アドオンモジュール

最新バージョンから新たに下記のアドオンモジュールが追加された。

6.1 BayoNet-Pro

独立行政法人産業技術総合研究所で開発された Bayesian Network 構築支援プログラム BayoNet を改良・機能追加し、VMStudio のアドオンモジュールとして BayoNet-Pro を開発した。BayoNet-Pro では以下の機能が実装されている。

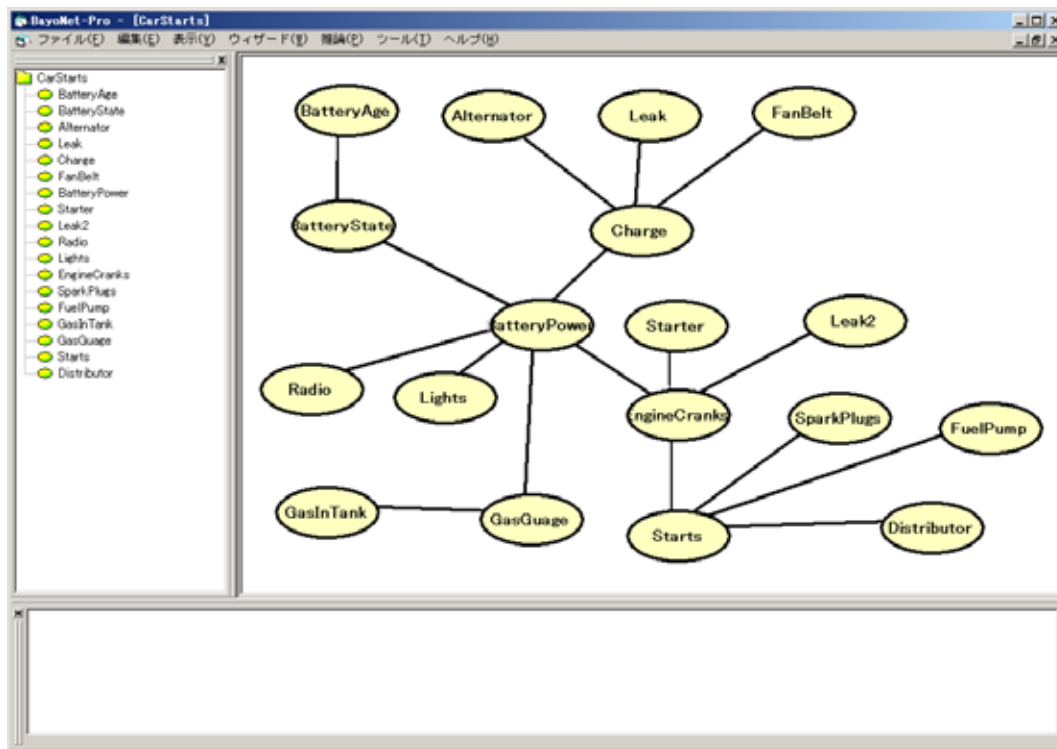
構築機能

Bayesian Network を、1) データから学習させ構築すること、2) ユーザが手で新規に作成したり、作成した Network を編集することが可能である。

推論機能

Bayesian Network を用いて、状態が既知のノード情報から、状態が未知のノードが、

ある状態をとる確率を計算する。推論の結果、最も確率の高い状態をそのノードの状態とみなすことで、分類分析での予測に対応する処理も行える。



6.2 テキストマイニング

テキストデータを解析するツールである。詳細は、別発表に参考されたい。

7.まとめ

VMStudio はビジュアルプログラミング環境を提供するだけでなく、データの加工処理から探索処理までの、データマイニングの全工程における広範なマイニングツールを提供する。また、汎用データ解析ソフト S-PLUS とシームレスに接続し、S-PLUS の全機能を使用することが出来る。VMStudio では、今後、データマイニングの最新技術をいち早く取り入れるだけでなく、マイニングの手法を組み合わせ、業界に特化したアドオンモジュールを開発していく予定である。例えば、バイオ関連のアドオンモジュールでは、DNA のマイクロアレイ等の、遺伝子(変数)の数が数千から数万になるようなデータから、病気に関係ある遺伝子(変数)を抽出するモデルを構築する。その他にも、流通業界、金融・保険業界、医療・薬品業界、製造業に特化したモジュールの開発も予定している。

==Visual Mining Studio Ver.3.0 新機能一覧==

1. ビジュアルプログラミング機能

- ・ループ処理機能
- ・ユーザ管理機能
- ・システムパラメータの追加

2. データ操作機能

- ・グルーピング機能
簡単な GUI 操作で数値データをカテゴリ化したり、カテゴリデータを再カテゴリ化
- ・カテゴリデータの数量化機能
列属性変更機能にカテゴリデータを数量化する機能を追加
- ・ソート機能 改良
ソートキー列毎に昇順・降順を指定可能に
- ・スクリプト編集機能 改良

3. データ探索機能

3.1 クラスタ分析

- ・自己組織化マップ
- ・カテゴリデータの取り扱い
クラスタ分析手法(K-Means、BIRCH、OPTICS、自己組織化マップ)でカテゴリデータも取り扱い可能に。

3.2 分類分析

- ・k-NN 分析 改良
回帰分機器が可能に
- ・Support Vector Machine 改良
多群判別が可能に
- ・Decision Tree 改良
- ・重み付学習機能
重みを付けて学習することが可能に。重みの指定は、行毎、クラス毎を指定可能
- ・モデル保存機能
作成したモデルを保存し、アイコンとして登録することが可能
- ・モデル統合機能
複数のモデルを統合する集団学習機能

3.3 アソシエーション分析

- ・ アソシエーション分析
高速化
- ・ 時系列アソシエーション分析
時系列情報を持ったデータのアソシエーション分析を行う機能
- ・ 関連性ダイアグラム分析 改良

4. アドオンモジュール

- ・ BayeoNet-Pro
Bayesian Network を構築し、推論を行う機能
- ・ テキストマイニング
テキストデータをマイニングする機能