共同研究全般にわたるデータベースの構築及び管理
４の３（地域分）

データマイニングツールの開発

株式会社数理システム（共同研究員）　徐　良為
（科学技術部長）　水田千益

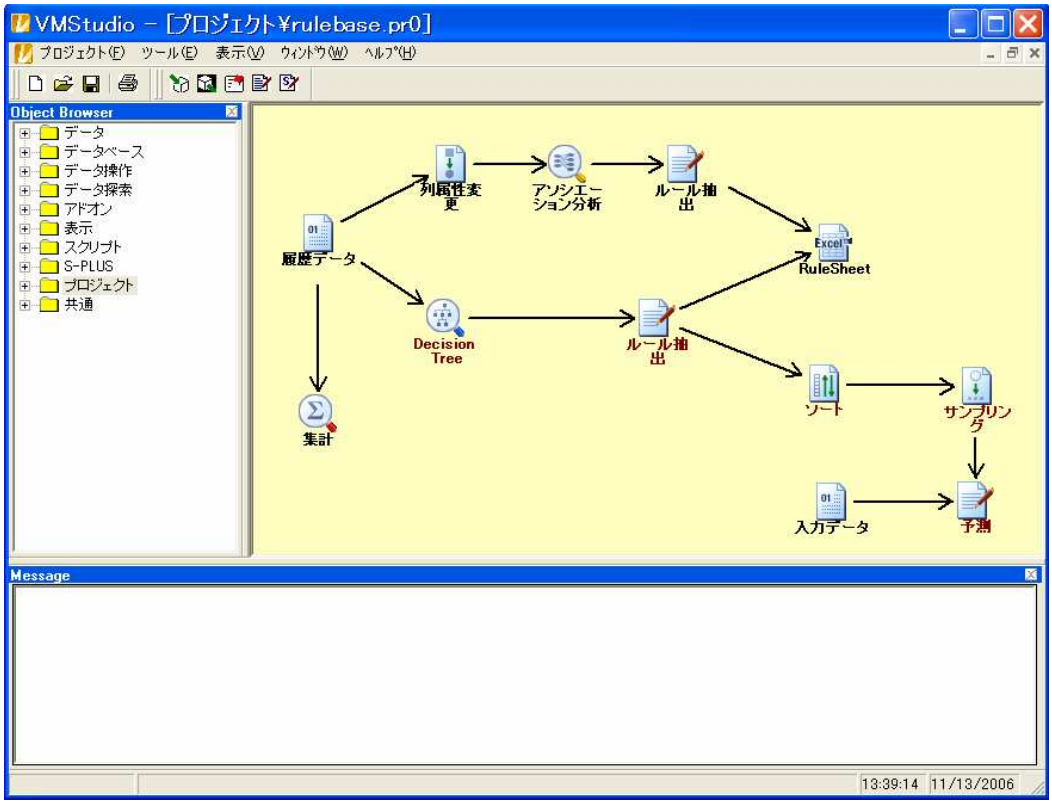
【目的と概要】

かずさDNA研究所においては、その保持するマウス長鎖cDNAライブラリーをもとに、マイクロアレイを利用して、遺伝子発現情報を収集することになっている。本収集データは、構築予定のマイクロアレイ・データベースシステムに格納される。そのデータは膨大なものになり、その解析には、規格化、フィルタリング、クラスタリングなどのデータマイニングの諸手法が必要となる。また、マイクロアレイ解析に特有な手法も必要である。本テーマの研究では、数理システム開発の汎用データマイニングツールを用い、マイクロアレイデータの解析に必要な手法の確立、およびそれをもとにしたマイクロアレイ専用データマイニングツールの開発を目指して来た。

【研究成果の概要と今後の取り込み】

（1）　統計解析及びデータマイニングの統合環境構築

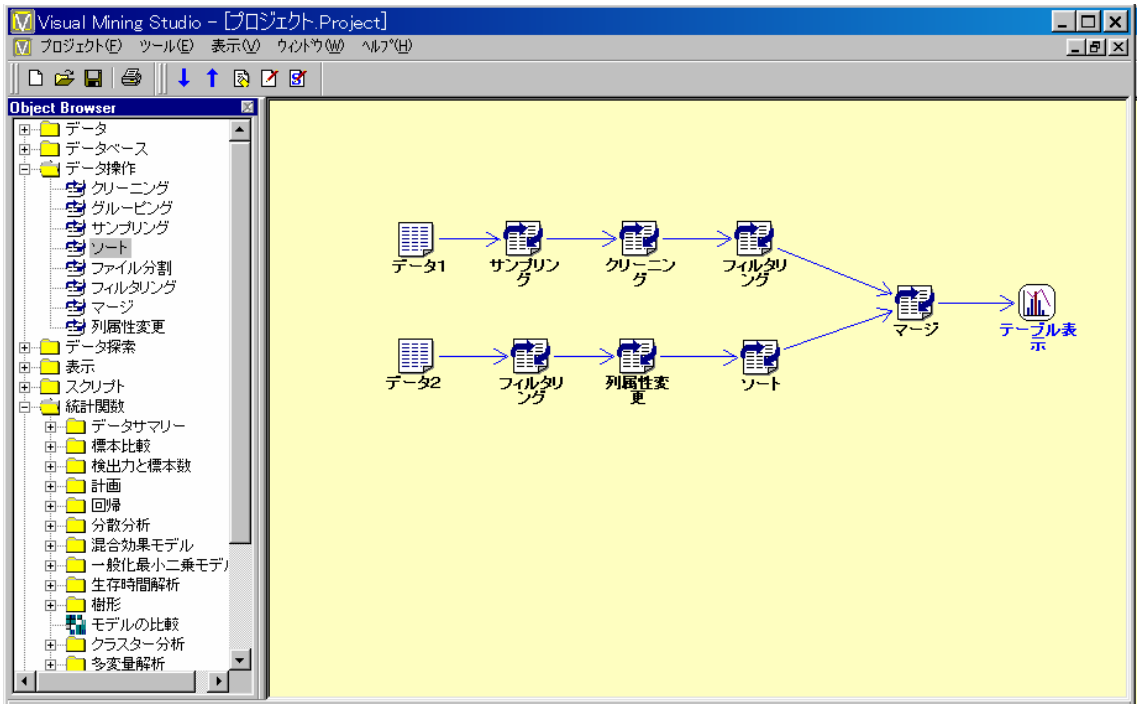
研究者が統計解析・データマイニングを行うプロセスを総合的に支援する環境を構築した。本システムユーザがGUI画面を通して、分析処理の流れを容易に行い、処理結果を確認しながら、トライ＆エラー分析プロジェクトを行うことが可能である。



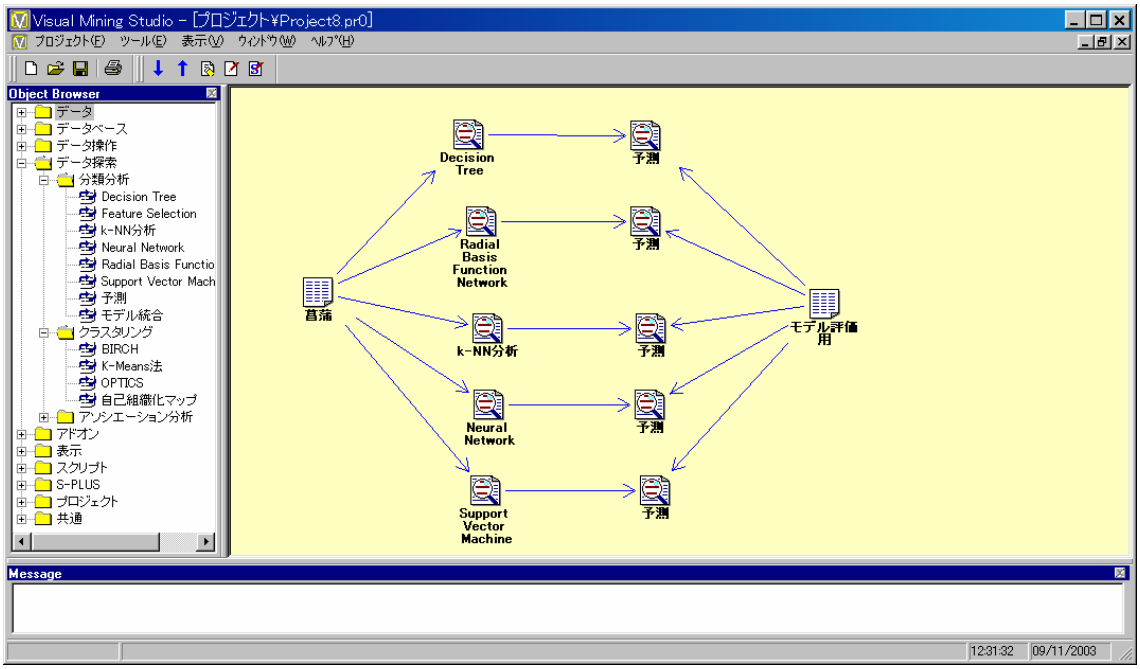
（2） マイニング最新手法の取り入れ

本システムは、本来汎用データマイニングツールとして設計、開発されたものであった。その汎用性を保ちながら、マイクロアレイ分析に特有な性質を考慮に入れることにし、ツールの修正及び新規追加を行った。

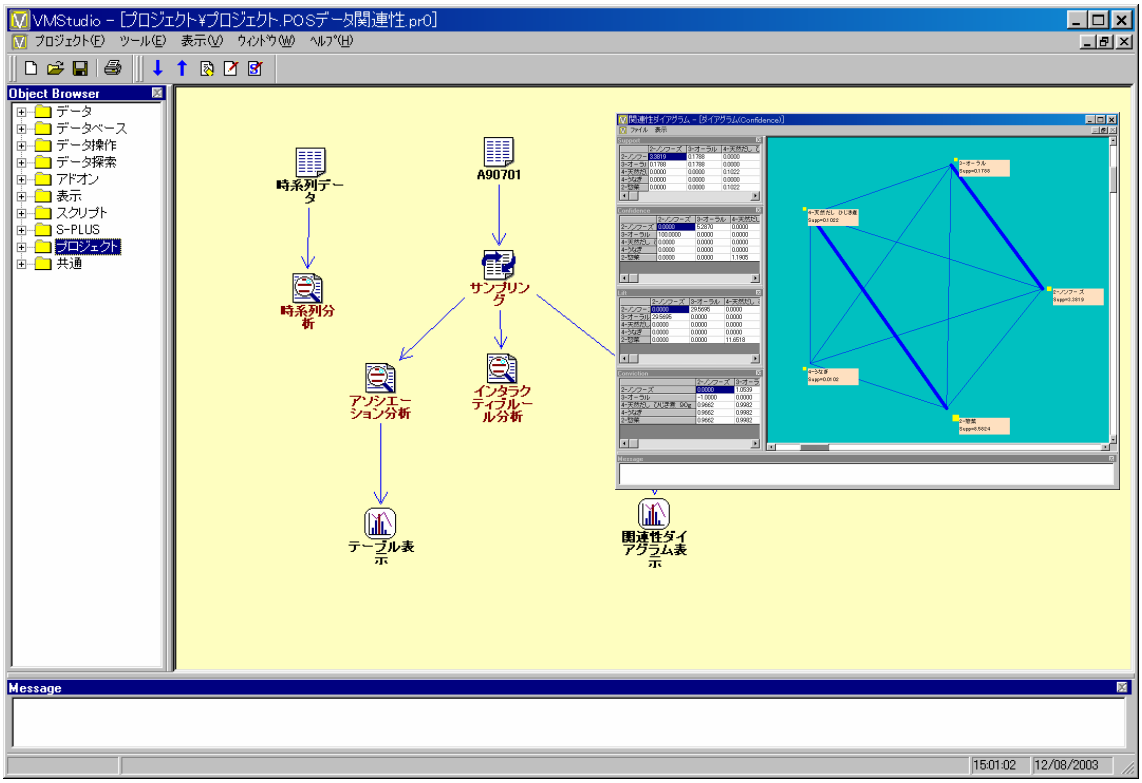
- データマイニング・分析の前処理
  本システムは、主にテーブル形式のデータ形式を対象としています。マイクロアレイを用いて、cDNA ライブラリに保持されるマウス長鎖cDNAライブラリから得られる遺伝子発現情報をテーブル形式にまとめ、不要情報の削除（変数選択）、欠損値補填、規格化、基本情報の集計などを試行錯誤しながら行いました。



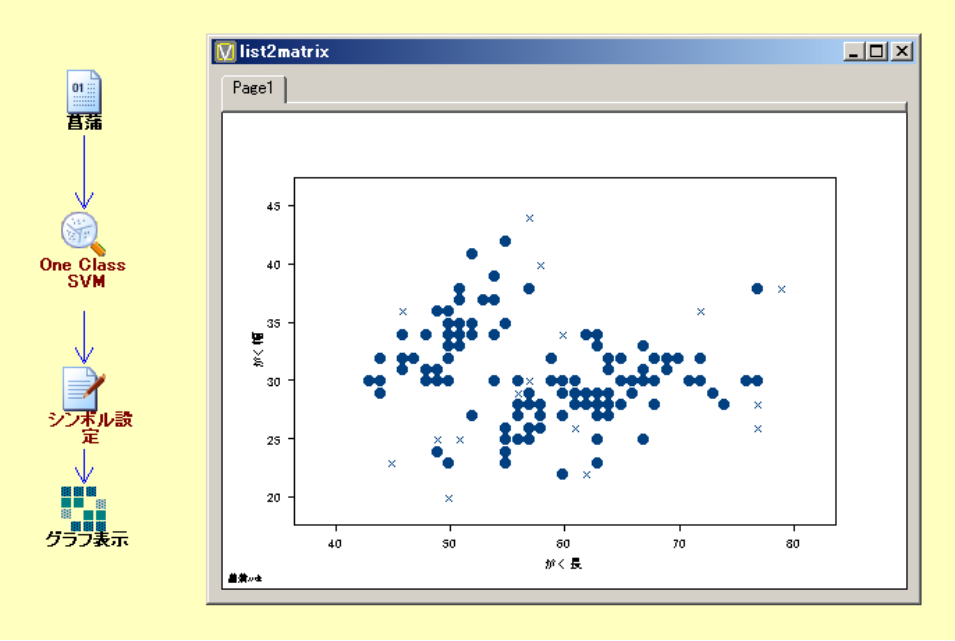
- 分類モデル、クラスタリング、特に　決定木、ニューラルネットワーク、K-NN、K-Means、BIRCH、OPTICS　について検討した



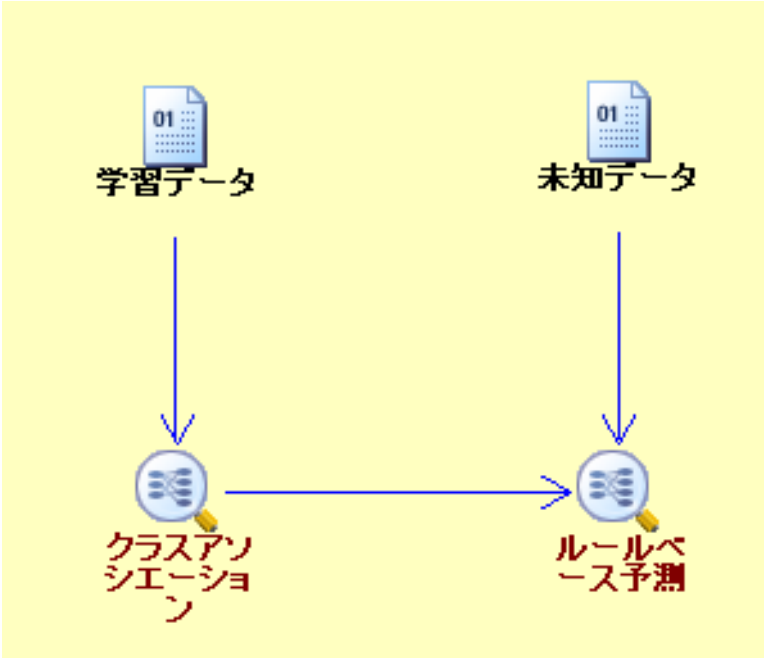
- アソシエーション分析：多階層、時系列、利便性、可視性の観点から実装、テストを行った。



- 分類ツール：Support Vector Machine の設計、実装。
- 分類ツール：Naïve Bayes の設計、実装。
- 外れ値の検出に有効と思われる　One Class SVM の設計、実装



・　クラスアソシエーション分析（特定目的変数の分類モデル）
・　ルールベース予測ツール



（3）　適用事例での検証

本システムがバイオ関連分析の有効性を検証するために、本システムを用いて、2001年のデータマイニングコンテストデータKDD-2001のテーマにして、次の２つ課題について検証した。

本検証を通じ、本システムはバイオデータにも有効であることを証明し、今回のコンテスト参加者が分析に利用したツール一覧は、ほぼ、本システムに含まれていることを判明した。

（4）　今後の取込み

土台が完成しつつあるだが、より多くマイクロアレイ分析テーマを検証し、バイオデータ分析に特化した分析ツールを作成して行きたい。