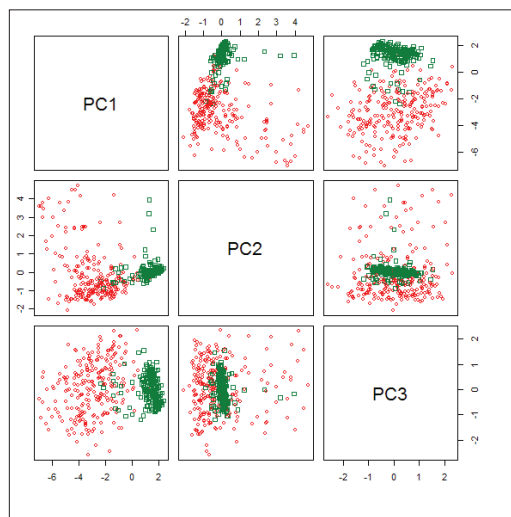


R STATISTICS COMPONENT COLLECTION

Pipeline PilotのR Statistics Component Collectionを使用すると、統計的な解析や結果のグラフィック表示を行い、その結果に基づいた意思決定を行えるようになります。このコレクションには、Rによるクラスタリング等のデータ解析や、統計モデル構築などの統計手法を実装したコンポーネントが含まれています。基盤となる統計エンジンは、広く使用されているオープンソースパッケージであるRです。このコレクションを使用すると、Rの統計分析とデータ操作のメソッドを Pipeline Pilot のデータストリームに適用できます。ユーザーは、Rの出力結果をパイプラインに直接取り込み、Pipeline Pilot フレームワークのコンポーネントを使用して詳細に分析することができます。既存のR スクリプトを Pipeline Pilot のカスタムコンポーネントで使用できるため、別のプロトコルで再利用したり、組織内で共有することができます。

R STATISTICS COLLECTIONを 使用すると、次のことが実現できます。

- ・ ヒートマップで値の相関を表示し、最も関連性の高いものを検出する
- ・ Box Plotを使用してデータの分布を表示する
- ・ 分散分析(ANOVA)を実行して複数のデータセット間の平均値に見られる差異を特定する
- ・ ロジスティック回帰分析、サポートベクターマシン(SVM)、ニューラルネットワーク、その他10種類の学習メソッドのいずれかを使用してデータをモデル化する
- ・ モデルが適切に適用されるよう支援する Model Applicability Domain(MAD)をサポートし、構築したモデルを適用して新しいデータセットを予測する
- ・ 訓練データを任意のモデルに保存し、実験データの追加に伴って拡張できるようにする
- ・ 多様なクラスタリング手法を適用する



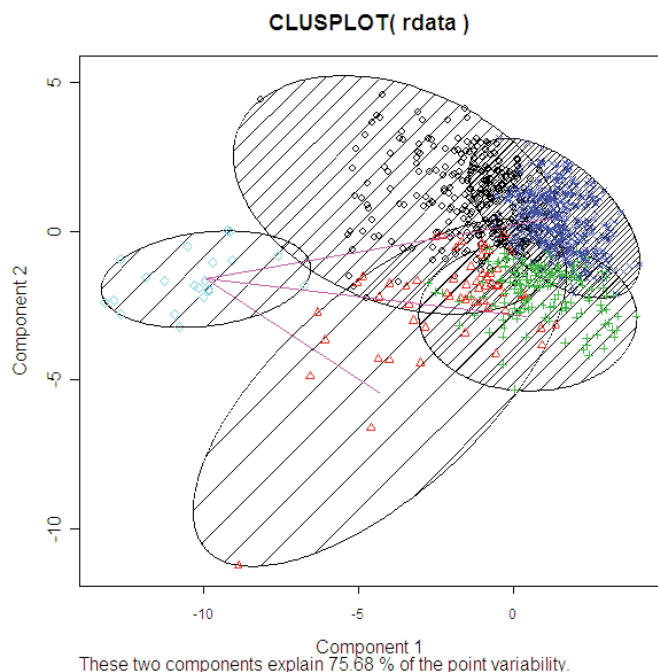
乳がんデータの主成分分析結果を組み合わせで表示

- ・ 独自の R スクリプトを Pipeline Pilot の個別のデータレコードやデータストリーム全体に適用する

分析

複数の測定結果を比較するためにt検定やANOVAを使用して有意性分析を行い、異なる測定結果の平均が同じであるかどうかを確認できます。R Correlation Matrixコンポーネントでは、一連の記述子に関するマトリックスを生成して記述子の相関度を確認したり、ヒートマップを作成して記述子空間におけるパターンを視覚的に確認することができます。このコンポーネントには次のものが含まれています。

- R ANOVA
- R K-Nearest Neighbors
- R Correlation Matrix
- R Principal Components Analysis
- R Probability Distributions
- R One-variable Tests
- R Factor Analysis
- R Two-variables Tests



Cluster Plot

クラスタリング

R を使用することで、あらゆる種類の Pipeline Pilot データと併用できる様々なクラスタリング手法を提供します。たとえば、フィンガープリントをRの階層的クラスタリング手法やk平均クラスタリング手法の記述子として分子データセットに使用することもできます。このコンポーネントには次のものが含まれています。

- R Cluster Agnes
- R Cluster Clara
- R Cluster Diana
- R Cluster Fanny
- R Cluster PAM
- R Cluster K-Means

データ操作

データセットが不完全だったり、不要な情報が含まれていたり、その他の理由により不規則である場合、Rデータ操作コンポーネントを使用して欠落値を置き換えたりデータをスムージングすることができます。このコンポーネントには次のものが含まれています。

- R Missing Values
- R Loess Smoother
- R Remove Zero-Variance Properties
- R Spline Smoother
- R Smoother
- R Friedman SuperSmoother

チャート

チャートは統計結果の分析や報告において非常に重要な役割を果たします。このコンポーネントでは、HTMLビューアに表示したりレポートに表示したりレポートに埋め込むことができるPNG画像を作成します。このコンポーネントには次のものが含まれています。

- R 2D Plot
- R 3D Plot
- R Pairs Plot
- R Conditional Plot
- R Histogram
- R Parallel Coordinates Plot
- R XY Plot

学習モデル

Pipeline PilotのData ModelingやAdvanced Data Modeling Collectionsの学習手法を補完するために、R Statistics Collectionでは、ニューラルネットワーク、サポートベクターマシン(SVM)のほか、数種類の統計的学習理論のモデルタイプを追加しました。分類の問題と回帰の問題に対する様々な手法がサポートされています。このコンポーネントには次のものが含まれています。

- Learn R Linear Model
- Learn R Linear Discriminant Analysis Model
- Learn R Generalized Linear Model
- Learn R Neural Net Model
- Learn R Non-Linear Model
- Learn R Support Vector Machine Model
- Learn R Logistic Regression Model
- Learn R Partial Least Squares Model

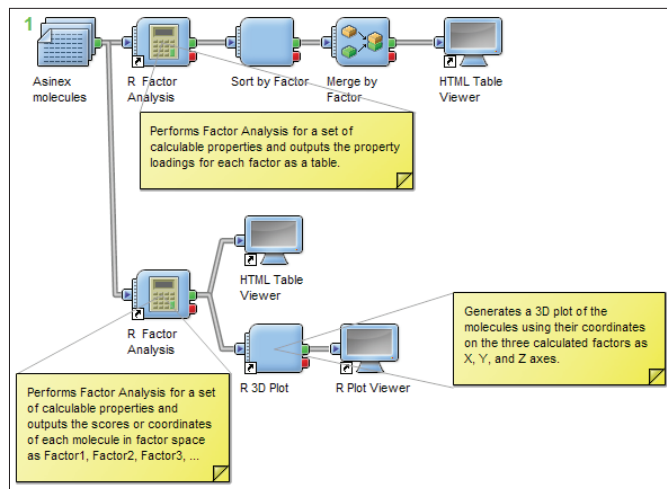
多次元スケーリング

データセットの特性は、データアイテム間のペアワイズ距離を含むマトリックスで示すことができます。R Statistics Collectionのコンポーネントでは、元の距離を可能な限り保持しながらデータを低次元空間に配置する多次元スケーリングを提供しています。このコンポーネントには次のものが含まれています。

- R Classical MDS
- R Sammon
- R Nonmetric MDS
- R Self Organizing Map

カスタマイズ

R Statistics Collection のコンポーネントはすべて、サブプロトコルとして実装されます。Rスクリプトを記述できるのであれば、コンポーネントの変更やカスタマイズを行って、さらに多くのRの機能をPipelinePilotに取り入れることができます。また、次の



Asinexデータの因子分析を示すプロトコルをR Statistics Component Collectionで構築

2つのコンポーネントを使用すれば、PipelinePilotのデータストリーム全体に、またはコンポーネントに入力される各データレコードに対して、Rスクリプトを適用できます。

- R Custom Script
- R Custom Script for Each Data

PIPELINE PILOTの概要

PipelinePilotは、さまざまな場所に保存されているデータから科学的価値を引き出し、科学的ワークフローを自動化して、より広範な科学コミュニティでのコラボレーションを促進することにより、研究開発組織の技術革新を支援する、拡張性に富んだ大規模サイエンティフィック・インフォマティクス・プラットフォームです。PipelinePilotのコンポーネントコレクションはプラットフォームの科学的な構成要素あり、科学的なカテゴリや機能でグループ化されています。コンポーネントをグラフィカルに組み合わせることで、データの取得、フィルタリング、分析レポート作成のワークフローを作成できます。

Pipeline Pilotの詳細については、次のURLを参照して下さい。

<http://accelrys.co.jp/products/pipeline-pilot>