

ADVANCED DATA MODELING COMPONENT COLLECTION

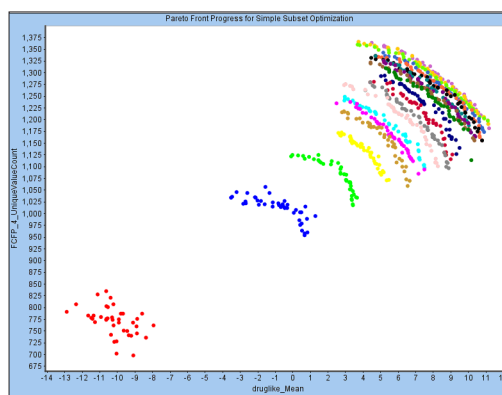
The Advanced Data Modeling Component Collection for Pipeline Pilot provides components for Recursive Partitioning (RP) classification models, Genetic Function Approximation (GFA) regression models, and multi-objective Pareto Optimization. The RP components provide multiple methods for building single tree or forest models, for either single or multiple response properties. The GFA components use a sophisticated genetic algorithm to perform variable selection and build multiple models, which can be combined into a consensus or ensemble model for more accurate predictions. The Pareto Optimization components provide methods for multi-objective optimization problems that provide solutions giving the best tradeoff among two or more conflicting goals.

WITH THE RECURSIVE PARTITIONING COMPONENTS YOU CAN:

- Perform rapid learning and data mining experiments on large data sets with large numbers of descriptors, including molecular data sets using fingerprints as descriptors
- Visualize trees to understand the relationships between descriptors and responses
- Analyze variable importance to identify the most discriminating descriptors
- Rapidly apply models to make predictions for new data sets, including model applicability domain (MAD) support to ensure the model is applied properly (also applies to GFA models)

WITH THE GFA COMPONENTS YOU CAN:

- Return multiple models rather than a single "best" model by creating a number of trial models, thereby generating multiple hypotheses for further investigation



Pareto front progress for simple subset optimization

- Combine multiple models into a single ensemble model, which often yields better predictive performance than any one of its component models
- Plot variable usage statistics over the evolution of the model population, giving insight into the descriptors most responsible for determining the response

WITH THE PARETO OPTIMIZATION COMPONENTS YOU CAN:

- Optimize solutions for problems as diverse as combinatorial library design, formulation ingredient optimization, or stock portfolio risk management
- Find individual samples within a data set that have the best tradeoff among desired property values
- Find subsets of samples from a larger data set that collectively have the best tradeoffs among desired property values

LEARNERS

The collection includes RP components to train single tree models, cross-validated single tree models and forest models. Parameters within the components provide extensive control over the learning method, allowing the selection of small forests or large forests of random trees as well as the size, depth, splitting, pruning, and weighting methods to be applied to the trees. The components allow the specification of single or multiple response (Y) variables to be modeled. The GFA components can create general property models, models of molecular properties, or mixture models of formulation data. All learners provide model applicability domain (MAD) support to help ensure that models

are applied properly when making predictions. Training data can be saved with any model, allowing the model to be extended as additional experimental data become available.

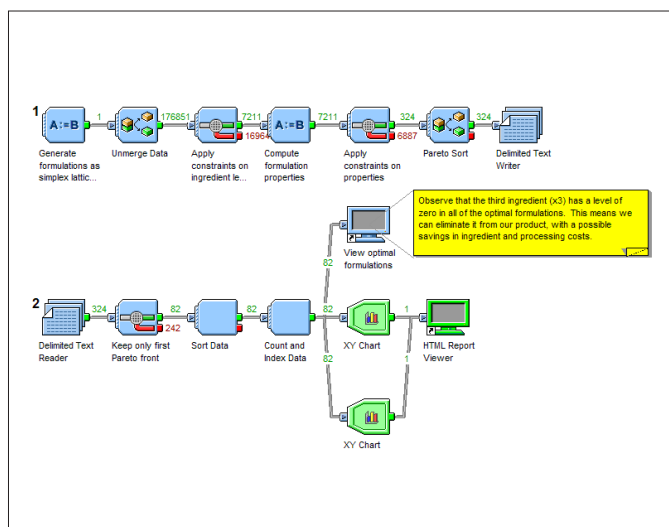
VIEWER

The Tree Model viewer provides an interactive, web-based visualization of one or more trees. You can browse and navigate to trees within a forest and within parts of a specific tree. The tree display shows the descriptors in use (including graphical display of molecular fingerprint fragments), the proportion of observations within each class and the splitting rules. You can click on a tree node to see samples assigned to that node and the rules that lead to that node. These rules are automatically converted into PilotScript that you can paste into a Custom Filter component to identify other data records that would satisfy the same rules.

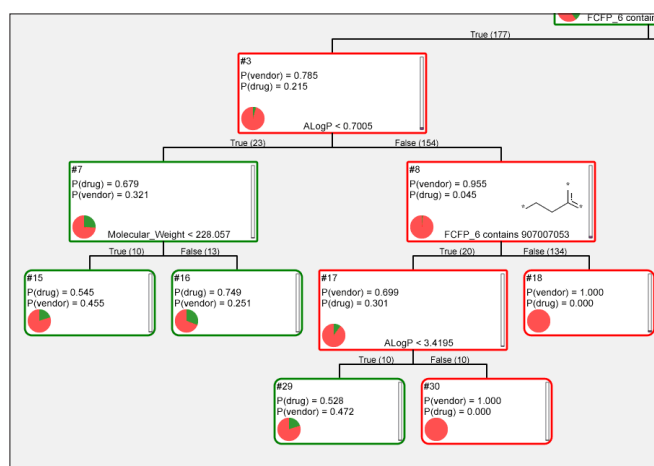
OPTIMIZERS

Components are provided for performing Pareto sorting, Pareto subset optimization, and Pareto combinatorial library optimization.

Pareto sorting ranks observations according to their Pareto score. You define the criteria that contribute to the score in terms of properties and goals for each property.



A protocol to optimize formulations with constraints on ingredient levels and properties



Visualize decision trees to understand the relationships between descriptors and responses

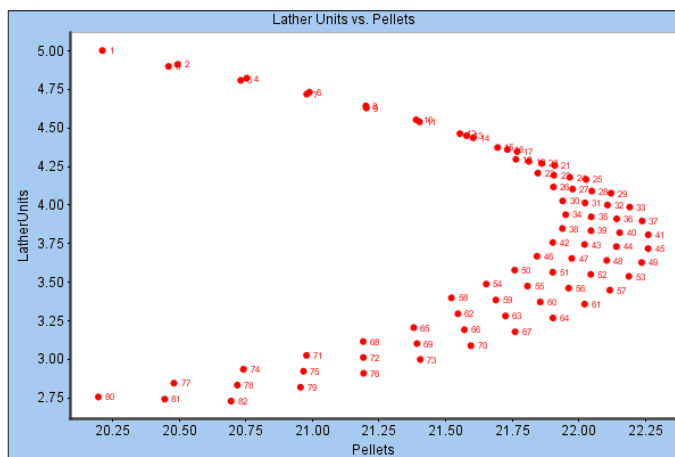
The Pareto subset optimizer provides a set of subset selections, each of which is on the Pareto front defining the best possible tradeoff among the goals you have selected.

The combinatorial optimizer includes a combinatorial constraint, so that rather than selecting a subset of samples, the optimizer selects reactants or groups that combine to produce the products. As an example in combinatorial chemical library design, you can specify the selection of a maximally diverse, maximally drug-like 8x12x20 combinatorial library from a larger library of 100x100x100 possible reactants.

ABOUT PIPELINE PILOT

Pipeline Pilot is an enterprise-scalable scientific informatics platform that enhances research and development organizations' ability to innovate by uncovering scientific value locked in disparate data silos, automating scientific workflows, and facilitating collaboration throughout the wider scientific community. Pipeline Pilot's Component Collections are the "scientific building blocks" of the platform and are grouped by category of science or function. By graphically combining components, you can construct workflows for data retrieval, filtering, analysis, and reporting.

To learn more about Pipeline Pilot, go to accelrys.com/pipeline-pilot



Optimized formulations with constraints on ingredient levels and properties