



Pentaho Data Mining

Copyright © 2009 Pentaho Corporation. Redistribution permitted. All trademarks are the property of their respective owners.

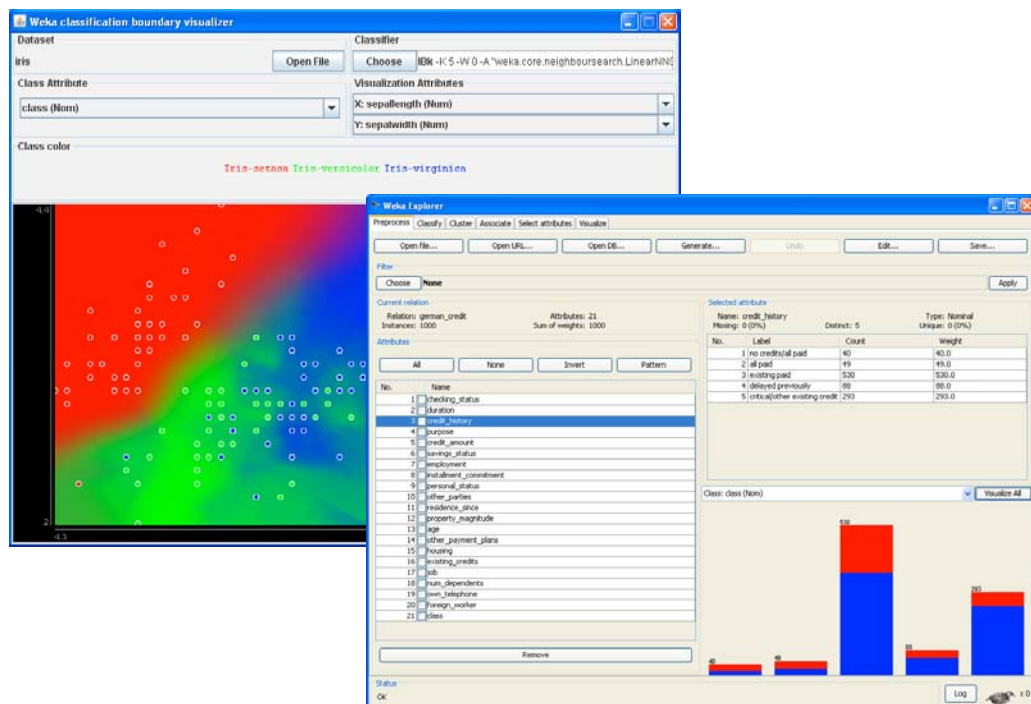
For the latest information, please visit our web site at www.pentaho.com

Pentaho Data Mining Overview

Once you've got analysis, reporting, and dashboards deployed, it's time to take your Business Intelligence (BI) to the next level by adding data mining and workflow to the mix. This is a level of BI excellence that many organizations never manage to evolve to, however the importance of pushing ahead with advanced capabilities cannot be underestimated – they can provide a truly sustainable competitive advantage and enable your organization to maximize both its efficiency and effectiveness.

Data Mining is the process of running data through sophisticated algorithms to uncover meaningful patterns and correlations that may otherwise be hidden. These can be used to help you understand the business better and also exploited to improve future performance through predictive analytics. For example, data mining can warn you there's a high probability a specific customer won't pay on time based on an analysis of customers with similar characteristics.

To help you fully utilize data mining for organizational advantage, the Pentaho BI Project team has worked in conjunction with the development and business communities to integrate mainstream BI capabilities with advanced data mining. Pentaho Data Mining is differentiated by its open, standards-compliant nature, use of the Weka data mining project, and tight integration with core business intelligence capabilities including data integration, reporting, analysis and dashboards. Other data mining offerings lack this level of sophistication and integration.



Pentaho Data Mining Provides Sophisticated Analytical Insight Into Trends and Opportunities.

Pentaho Data Mining Enterprise Edition

Pentaho Data Mining Enterprise Edition extends Pentaho's best-in-class open source business intelligence (BI) capabilities with additional software and services designed to help you and your organization:

- **Achieve BI success**
- **Save time, resources, and money**
- **Mitigate risk**

Achieve BI success

What makes the difference between success and failure in business intelligence or data warehousing projects? There is ample evidence from IT professionals, consultants, and industry analysts that success or failure with business intelligence is often driven far more by "people and process" issues rather than technology. Poor planning, lack of commitment, inadequate resources or skill sets, and inability to deliver initial results quickly can doom a BI project regardless of the selected software products and technology. While open source software is rapidly transforming the IT landscape and has provided new levels of flexibility and freedom for customers, open source software alone does not address the traditional pitfalls of BI projects. Pentaho Data Mining Enterprise Edition provides the product capabilities and value-added services to help you deliver a successful BI project for your organization, including consultative support and product expertise, software maintenance, software assurance, and more.

Save Time, Resources, and Money

Even large organizations have fewer IT resources than they would like, and they strive to get the most out of their investments in time, people, and technology. There are numerous public examples of Pentaho customers who have realized the Total Cost of Ownership (TCO) advantages of commercial open source BI from Pentaho, recognizing that investing in a relationship with Pentaho saves time, resources, and money not just in the long-term, but in the *short term* as they initiate BI projects. "Going it alone" with free BI software not only increases your risk of failure, it turns out to be more expensive. Pentaho Data Mining Enterprise Edition delivers critical benefits like stabilized software, direct access to product expertise, and committed response times to help you save time, resources, and money.

Mitigate Risk

Business intelligence risk comes in many shapes and forms. Risk of project failure, risk of late delivery, risk of going over budget, and legal risk as well. Beyond providing the software enhancements and services to reduce project risk, Pentaho provides a lower-cost model for enterprise-class business intelligence software that reduces budget risk by eliminating large, up-front software license fees. Pentaho Data Mining Enterprise Edition also includes legal protection to minimize your company's risk and exposure to potential legal issues related to intellectual property in open source software.

Pentaho Data Mining Enterprise Edition Features

Pentaho Data Mining Enterprise Edition allows you to deploy in production with confidence, security, and far lower total cost of ownership than proprietary alternatives. Pentaho Data Mining Enterprise Edition provides additional capabilities including comprehensive professional technical support, software maintenance, certified software, product expertise, and the best software assurance program in the industry.



Software and Services	Community Edition	Enterprise Edition
Data Mining	Open Source	Certified
Business Intelligence Platform	Open Source	Certified
Community Forums Interaction	✓	✓
Community Web Documentation (wiki)	✓	✓
Professional Support		
• Telephone support (toll-free)		✓
• E-mail support		✓
• Service Level Agreement		✓
• Unlimited support cases		✓
Software Maintenance		
• Software maintenance	By in-house staff	✓ By Pentaho Engineers
• Patch releases		✓
• Fixes included in future releases		✓
Certified Software		
• Stabilized software		✓
• Managed release cycle		✓
• Optimized builds		✓
Product Expertise		

• Professional documentation		✓
• Knowledge base		✓
• Consultative support		✓
• Remote assistance packages		✓ Optional Add-On
• Installation/configuration packages		✓
• Design and integration packages		✓
• Troubleshooting and optimization packages		✓
• Enterprise Edition online forum		✓
• Web based training		✓ Optional Add-On
Software Assurance		
• Intellectual Property Indemnification		✓
• Warranty for services		✓

For more information on the features and benefits of Pentaho's Enterprise Editions, please see the [Pentaho BI Suite Enterprise Edition brochure](#).

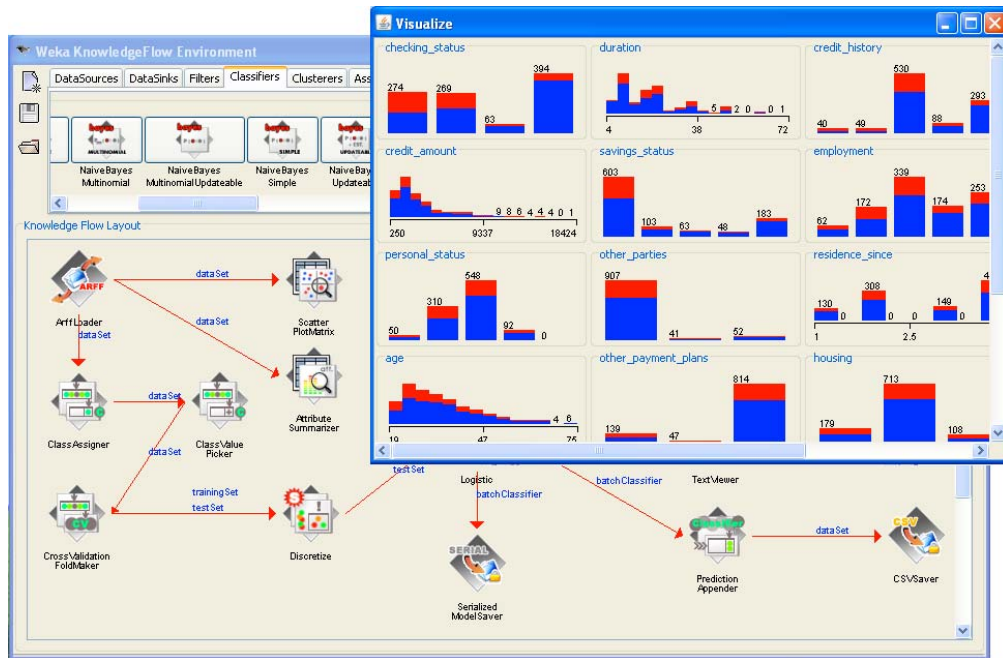
Pentaho Data Mining Feature Details

Powerful Data Mining Engine

- Provides a comprehensive set of machine learning algorithms from the Weka project including clustering, segmentation, decision trees, random forests, neural networks, and principal component analysis.
- Pentaho has added integration with the Pentaho Data Integration and automated the process of transforming data into the format the data mining engine needs.
- Algorithms can either be applied directly to a dataset or called from Java code.
- Output can be viewed graphically, interacted with programmatically, or used data source for reports, further analysis, and other processes.
- Filters are provided for discretization, normalization, re-sampling, attribute selection, and transforming and combining attributes.
- Classifiers provide models for predicting nominal or numeric quantities. Learning schemes include decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, Bayes' nets, and other advanced techniques.
- The data mining engine is also well-suited for developing new machine learning schemes, enabling customers to incorporate their own models.
- Inputs and outputs can be controlled programmatically, enabling developers to create completely custom solutions using the components provided.
- Support for Predictive Model Markup Language (PMML)

Graphical Design Tools

Graphical user interfaces are provided for data pre-processing, classification, regression, clustering, association rules, and visualization.



KnowledgeFlow shows you the flow of data through the system and the processes that it goes through.

Uncover Hidden Patterns and Relationships

A classic example of data mining is a retailer who uncovers a relationship between sales of diapers and beer on Sunday afternoons – two items you wouldn't normally consider as linked. The explanation is that husbands who are sent out to pick up a fresh supply of diapers are also likely to pick up some beer while they happen to be in the store – something that hadn't been recognized as a significant sales driver before data mining uncovered it.

Exploit Insights to Improve Performance

Continuing the example above, very often retailers act on the relationships they discover by using tactics such as placing linked items together on end-of-isle displays as a way to spur additional purchases. All organizations can benefit from acting in a similar way – using newly discovered patterns and correlations as the basis for taking action to improve their efficiency and effectiveness.

Predict Future Performance

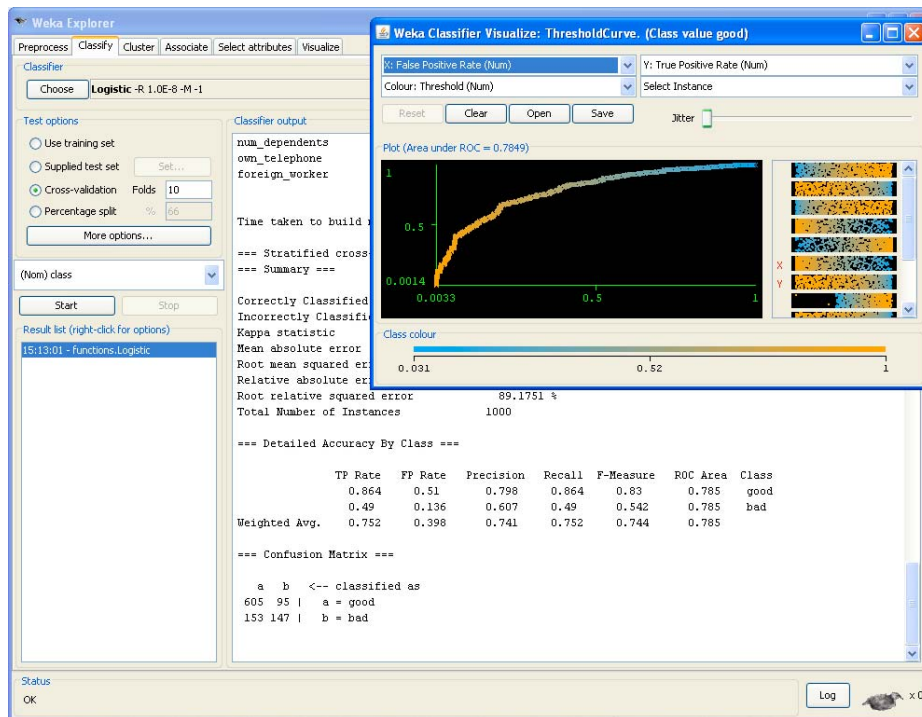
"Those who do not learn from history are doomed to repeat it" is a famous quote from philosopher George Santayana. In the case of data mining, being able to predict outcomes based on historic data can dramatically improve the quality and outcomes of decision making in the present. As a simple example, if the best indicator of whether a customer will pay on time turns out to be a combination of their market segment and whether or not they have paid previous bills on time, then this is information you can usefully benefit from in making current credit decisions.

Embed Insights into Your Applications

You can use the data mining results to display a simple summary statement and recommendations within operational applications. For example, on a credit screen you could add: “Based on this new account profile there is an 85% chance this customer will pay late. It is therefore recommended you require a 50% prepayment on this order”. Reporting on aggregate results such as Days Sales Outstanding (DSO) enables you to measure business improvements based on when recommendations were followed and when they weren't so that you can fine-tune your model and recommendations over time for optimal effect.

Wide Range of Algorithms

No algorithm is likely to be optimal in all situations. For this reason it's important that you're able to try out a range to find the algorithm that fits a particular set of data the best. If you find several data mining algorithms that fit well, you can use all of them - for example: “Based on analysis of 3 predictive models, the chances this customer will pay late are; Model A: 95% (96% correct), Model B: 89% (92% correct), Model C: 76% (97% correct)”.



Knowledge Explorer lets you explore your data and prepare it for data mining.

How Data Mining Works

Choosing a Model

Analysts can work with a range of models graphically. These include many advanced forms of data mining such as clustering, segmentation, decision trees, random forests, neural nets, and principal component analysis.

Adding Data

Value-added features can be added to the data. For example, you can specify thresholds and have the system automatically “bucket” or derive data to create new columns for analysis.

Adapting

Each model works to adapt its parameters to attempt a best fit to the sample data. Analysts can let this happen automatically, or manually adjust parameters (depending on the model).

Evaluating

Results can be evaluated by applying the model to historical data to test its predictive power compared to actual results.

Perfecting

The cycle of adapting the model until it is optimized is known as “training” the model. Once properly trained, the model will reliably yield the best results for the specific business purpose it is being applied to.

Delivering

Output can be in a multitude of forms. For example, you might choose to include a simple statement within another application, or output a graphical decision tree that users can navigate.