

RAPID|MINER



RAPID|ANALYTICS

FACTSHEET

business analytics fast and powerful



rapid-i
REPORT THE FUTURE

RapidMiner and RapidAnalytics **RAPID|MINER AND RAPID|ANALYTICS**

BUSINESS ANALYTICS FAST AND POWERFUL

What are RapidMiner and RapidAnalytics?

RapidMiner is a complete business analytics workbench with a strong focus on data mining, text mining, and predictive analytics. It uses a wide variety of descriptive and predictive techniques to give you the insight to make profitable decisions. RapidAnalytics is an enterprise analytics server which also offers full reporting and dashboarding capabilities and, therefore, a complete business intelligence solution in combination with predictive analytics.

Why RapidMiner and RapidAnalytics?

- No software license fees
- Flexible/affordable support options
- Fastest development even of highly complex data mining processes
- Guaranteed operational reliability
- Most comprehensible and flexible data mining solution in the market
- Installation takes less than 5 min.
- Enterprise-ready performance and scalability for big data analytics
- Innovative analyst support

Company Facts

- Rapid-I offers software and services for business analytics
- Founded in 2006, headquartered in Dortmund, Germany
- Development since 2001
- More than 30 partners on all continents
- More than 3 million product downloads
- More than 35,000 production deployments
- More than 400 customers in more than 40 countries

Introduction

Rapid-I's flagship product RapidMiner and its server solution RapidAnalytics provide clients with industry leading business analytics tools and extensive benefits.

RAPID-I LOOKS INTO THE FUTURE.

When making decisions, our customers do not need merely rely on the gut feeling they get from looking at retrospective data. Instead, business analytics supplies the patterns hidden in large masses of structured and unstructured data using innovative techniques from the areas of predictive analytics, data mining and text mining. Rapid-I provides its customers with a profound insight into the most probable future. Rapid-I calls this: *Report the Future.*

WHAT DOES RAPID-I OFFER?

Rapid-I acts as a one-stop shop for all necessary software solutions and services for business analytics and continues to consistently develop this unique position in the open source environment with the help

of the active community. Thanks to its customer experience, Rapid-I leverages its rich range of successful business analytics applications within the market. Rapid-I was the first open source provider to explore innovative ranges of application together with its customers, including the areas of churn prevention, up and cross-selling, sales prediction, risk detection, fraud detection, predictive maintenance, price predictions and social media analysis.

WHAT ARE RAPIDMINER AND RAPIDANALYTICS?

The flagship, open source solution RapidMiner, has now been developed since 2001 and is one of the world's most-used solutions for data analyses today. The business analytics server RapidAnalytics uses RapidMiner as data processing engine for creating predictive and descriptive models and also offers full reporting and dashboarding capabilities. In addition, Rapid-I is currently the only open source provider to offer the complete range for 100% integrated business analytics, comparable to SAS or IBM in the proprietary environment. This begins with data warehousing and ETL and

ends with the creation of web-based reporting and dashboards. Unlike traditional solutions, the Rapid-I products follow a unique procedure, where each transformation, each visualization and each analysis is modeled as one process with the same tool. This process-based procedure of the Rapid-I software is unique for business analytics and offers the advantage that even the most complex of analyses and resulting predictions can be integrated directly into the infrastructure of the customer and his business processes. The complete covering of all relevant solution components and the integration based strictly on web services make Rapid-I's solutions stand out from those of our competitors.

FOR WHOM ARE RAPIDMINER AND RAPIDANALYTICS DESIGNED?

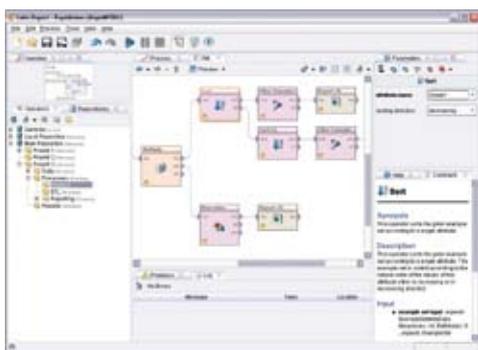
RapidMiner and its server solution RapidAnalytics already have a wide user community with over 400 customers in over 40 countries worldwide, including many small and medium-sized enterprises alongside global enterprises such as Sanofi, Miele, E-On, RWE, Tchibo, Libri, Pepsi, Lufthansa Systems, EADS, Telekom (German Telecom), LBB and GfK.



Key Benefits

AN ORGANIZED AND LOGICAL GUI FOR BUSINESS ANALYTICS SUCCESS.

Everything is a process! RapidMiner is a powerful and intuitive graphical user interface for the design of analysis processes. Compared to standard ETL tools for example, Rapid-I takes a more avant-garde approach: Each transformation, each visualisation, each analysis, each prediction and even each component of a web-based report is a process. This means one tool for all tasks. A further advantage: The processes can interact with one another and be transformed for integration (including into web services) with a



The intuitive GUI of Rapid-Miner allows the design of arbitrary data transformation and analysis processes. Such processes are self-documenting and can be updated and reused easily for new problems.

few clicks. Even complex dashboards with numerous element interactions can be created simply and without any programming. This paradigm shift is often unfamiliar to users of traditional BI software at first, but is soon very much appreciated by all users thanks to the higher flexibility, power, the smooth integration of ETL and analyses as well as its much greater ease of integration.

BIG DATA ANALYTICS MADE EASY.

The fundamental goal of data mining has always been to find connections in very large data volumes. In the past few years, Rapid-I established a big data initiative focused on the topic of big data and developed several techniques:

- *Stream mining:* Instead of holding complete data sets in the memory, only parts of the data are taken through an analysis process and the part results aggregated in a suitable location later on. Such part processes can also be carried out in distributed form, e.g. in RapidAnalytics clusters or even on Hadoop.
- *In-database-mining:* Instead of taking the data to the algorithm,

this extension supports taking the algorithms to the data. Thus the execution of analyses, and in particular of a scoring, is directly supported within databases. Until now, such a solution has only been available from individual database providers such as Oracle and IBM DB2 and on a very limited basis. Rapid-I now offers this solution for numerous analysis procedures and database-wide.

- *Radoop:* The world's first graphical connection of Hadoop for the handling of big data analytics, meaning that even terabytes and petabytes of data can be transformed and analysed. Radoop combines the strengths of Rapid-Miner with Hadoop. The result is a solution for the graphical development and execution of workflows for ETL and predictive analytics on Hadoop clusters including support for the Hadoop file system, Hive and Mahout.

ANALYSES WITH UNPRECEDENTED EASE AND USER SUPPORT.

No other tool on the market offers nearly as comprehensive an analyst support. Thanks to the administration of meta data and intelligent

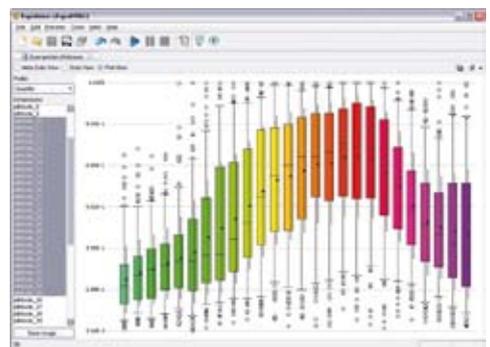
analyses of the process design itself, the user is assisted very closely during his work:

- *Meta data propagation*: No more trial and error but real-time inspection of the expected results as early as design time without having to wait for potentially lengthy process executions,
- *On-the-fly error detection*: If RapidMiner recognizes an error in the design of a process, potential problems are caught immediately during the design phase and not only during process execution,
- *Quick fixes*: If RapidMiner recognises an error in the design and can suggest one or more possible solutions, these are just a click away,
- *Profiler*: RapidMiner can continually observe the storage and runtime behaviour of analysis processes in the background and identify possible bottlenecks. Clear visualisations help to optimise the performance of the processes.
- *Community Extension*: This RapidMiner extension connects you with the Web 2.0 portal myExperiment.org sharing your processes with other analysts. You can then tag and document these

processes downloading them directly back into RapidMiner with one click.

- *Recommender*: RapidMiner continually analyses the analysis process making suggestions such as “analysts who created a similar process to you used step / operator X next” helping not only the beginner, but also accelerating the expert’s work tremendously,
- *Intelligent discovery assistant*: This assistant automatically creates analysis processes on the basis of an input data and analysis target specification. The assistant is based on the learnings from thousands of analysis processes and suggests analysis steps that are necessary for the given data set or promise particularly good results.
- *MLWizard*: This assistant concentrates solely on the generation of optimal classification processes for a given data set. For this purpose, MLWizard computes so-called land-marking features of the data and generates a prediction from this as to which procedure is likely to achieve the best performance.

RapidMiner offers numerous visualization and chart options. Additionally to the large amount number of data visualizations, all types of models like decision trees, cluster models, or attribute weights can also be visualized.



MORE POWER AND COLLABORATION WITH RAPIDANALYTICS.

RapidAnalytics is the first open source server for business analytics supporting the full range from ETL over data mining to interactive web-based reports and dashboards. RapidAnalytics uses RapidMiner as an engine and offers, among other things, the remote and scheduled execution of analysis processes, shared repositories for collaborative working, user management, web-based access to reports, dashboards, results and processes / web services. A special feature of RapidAnalytics is the ability to connect further Rapid-I products, including RapidNet for the visualisation and analysis of network structures.

AN INTEGRATED SUITE OF UNMATCHED ANALYTICAL TECHNIQUES.

Over 1,500 methods for data integration, data transformation, analysis and modelling as well as visualisation - no other solution on the market offers more procedures and more possibilities of defining the optimal analysis processes. RapidMiner and RapidAnalytics offer numerous procedures, especially in the area of attribute selection and for outlier detection, which no other solution offers.

DON'T LIKE SURPRISES? PERFORM FAIR AND CORRECT PERFORMANCE ESTIMATIONS.

RapidMiner and RapidAnalytics are the only solution for data mining



RapidMiner and RapidAnalytics and their extensions offer more than 1500 operations for all tasks in data transformation, analysis, and visualization. Popular extensions include a connectors to R and the machine learning library Weka, extensions for text and web mining as well as those for time series analyses.

which offer pre-processing models for all important types of data transformations. This is fundamental for two reasons. First, this approach ensures that the data in the scoring phase is transformed in exactly the same way as in the training phase. And second and more importantly, this will prevent slipping information from training to the model application.

For example, if the analyst decides for normalization before model training, the normalization factors derived from the testing data should not be used during training phase. Only RapidMiner's strict division between modelling and pre-processing into their own operators, instead of automatically performing the pre-processing inside of the modelling, together with the concept of pre-processing models makes sure that unseen information stays unseen. This is not only necessary for normalization or transformations but also for selecting attributes or applying other types of data dependent transformations. RapidMiner's unique approach is the only guarantee that no overfitting is introduced by pre-processing and a fair measurement of estimated performance can happen.

MARKETPLACE AND EXTENSIONS.

Rapid-I, as well as third-party providers and the community, offer numerous extensions for RapidMiner and RapidAnalytics. These are offered via the Rapid-I Marketplace, an app store for analytical solutions and algorithms. The Marketplace is the only one of its kind to date and serves for the distribution of further open source extensions and the commercial benefit of third-party providers' proprietary extensions. In addition to the extensions like Radoop or MLWizard discussed above, the following innovative extensions are available among the growing 50-plus others:

- *R connector*: Arbitrary R models and scripts can be seamlessly integrated into RapidMiner processes; the R perspective offers the known R console together with plotting facilities of R, online help, and multi-line editing with syntax highlighting.
- *Weka*: Adds more than 100 additional operators from the well-known machine learning library Weka.
- *Text*: Offers statistical text analysis by loading texts from different data sources including plain texts,

HTML, PDF, or RTF. Transform texts by a huge set of different filtering techniques including tokenization, stemming, stop word filtering, part of speech tagging, n-grams, dictionaries, and many more.

- **Web:** Access to internet sources including web pages, RSS feeds, and web services. Offers specific operators for handling the content of web pages.
- **Semantic web:** Enables the modeling of data from the semantic web. Supports the transformation of RDF triples into structured data sets and offers specific modelling techniques.
- **Image processing:** Enables the extraction of descriptive features from graphic data and specific data transformations as well as analyses, including segments and detection of image sections.
- **Information extraction:** Offers techniques for the extraction of information from unstructured texts including the determining of so-called named entities and of relations under these.



The server RapidAnalytics offers a complete web-based suite for the pixel-perfect creation of interactive reports and dashboards – using also data mining additionally to the mere visualizations of retrospective data.

DEVELOP? DEPLOY!

The Rapid-I products support numerous standards, including PMML. The PMML standard serves for exchanging predictive models with other systems, in particular databases, and thus for effectively applying the models on large data volumes (scoring). Alongside IBM, MicroStrategy, SAS, SPSS and others, Rapid-I is a member in the Data Mining Group, the consortium for the creation and continued development of the PMML standard. Rapid-I is also a founding member of the Data Mining Foundry, which aims at the standardization of a data mining ontology as well as an exchange

standard for analytical processes where RapidMiner's XML format defines the baseline for the standardization efforts.

Key Features

Multiple interfaces

- Easy-to-use GUI for building analytics processes:
 - Simple and fast design of better models
 - XML process exchange
 - Reuse processes as templates for other projects or users
 - Access Rapid-I scripting environment or the statistical language R from within your process
 - Every analysis is a process, each transformation or analysis step is an operator
 - More than 1000 operators for all tasks of data transformation and analysis
- Batch processing:
 - Encapsulates all features of the GUI
 - Can be controlled with macros / external variables
 - Embed processes for data transformation, training, and scoring into customized applications via command line, Java API, or web services

Accessing and managing data

- Access to more than 40 file types
- Wizards for Microsoft Excel, Microsoft Access, comma-delimited files, and database

- connections
- Support for all databases via JDBC or ODBC including Oracle, IBM DB2, Microsoft SQL Server, MySQL, Postgres, Teradata, Ingres, VectorWise, and many more
- Connector to SAP
- Access to text documents and web pages, PDF, HTML, and XML
- Access to time series data, audio files, images, and many more
- Repository-based data management on local file systems or central servers via RapidAnalytics
- Enhanced data and meta data editor for repository entries
- Select attributes operator
- Aggregations for multiple groups and functions like sum, average, median, standard deviation, variance, count, least, mode, minimum, maximum, product, or log product
- Set operators like join, merge, append, union, or intersect
- Operators for handling meta data like rename or attribute role definition
- Filtering rows / examples according to range, missing values, wrong or correct predictions, or specific attribute values
- Filtering outliers according to distances, densities, local outlier

- factors, class outlier factors, local correlation integrals, or clustering based outlier detections
- Identification and removal of duplicates

Model import and export

- Storing of models in central repositories for reuse in other processes and projects
- Support of PMML 3.2 and 4.0
- Rapid-I is member of the Data Mining Group (PMML) as well as of the Data Mining Foundry creating standards for data mining ontologies (DMO) and process definition standards
- Import and export of RapidMiner models, R models, and Weka models from repository or files

Scoring

- Operator for applying models to data sets (scoring)
- Support of predictive models as well as cluster models, preprocessing models, transformation models, and models for missing value imputations
- Applying a model creates optimal scores by ignoring unused attributes and handling previously unseen values
- In-database scoring possible for

many non-PMML models

Sampling

- Absolute, relative, or probability-based
- Balanced
- Stratified
- Bootstrapping
- Model-based
- Kennard-Stone
- Range

Data partitioning

- Create training, validation, and test data sets
- Ensure high model quality through hold-out data sets
- Default stratification by the class if available
- User-defined partitions possible
- Resulting in example sets usable for modeling or further transformations

Transformations

- Normalization and standardization
- Z-transformation, range transformation, proportion transformation, or interquartile ranges
- Preprocessing models for applying the same transformations on test / scoring data
- De-normalization making use of preprocessing models
- Scaling by weights
- Adjustment of calendar dates and times
- All kinds of type conversions between numerical attributes, nominal / categorical attributes, and date attributes
- Operator for guessing correct

meta data from existing data sets

- Sorting and Pareto sort
- Shuffling
- Rotations: Pivoting, De-Pivoting, and transposing data sets
- Expression builder for arbitrary transformations on attributes:
 - Basic functions: addition, subtraction, multiplication, division, less than, greater than, less or equal, greater or equal, equal, not equal, Boolean not, Boolean and, Boolean or
 - Log and exponential functions: natural logarithm, logarithm base 10, logarithm dualis, exponential, power
 - Trigonometric functions: sine, cosine, tangent, arc sine, arc cosine, arc tangent, hyperbolic sine, hyperbolic cosine, hyperbolic tangent, inverse hyperbolic sine, inverse hyperbolic cosine, inverse hyperbolic tangent
 - Statistical functions: round, floor, ceiling, average, minimum, maximum
 - Text functions: to string, to number, cut, concatenation, replace and replace all, lower, upper, index, length, character at, compare, contains, equals, starts with, ends with, matches, suffix, prefix, trim, escape HTML
 - Date functions: parse, parse with locale, parse custom, before, after, to string, to string with locale, to string with custom pattern, create current, difference, add, set, get
 - Miscellaneous functions: if-then-else, square root, signum, random, modulus, sum, binomial, missing

Binning

- Interactive binning by user specification
- Simple binning
- Count-based
- Size-based
- Frequency-based
- Entropy-based minimizing the entropy in the induced partitions
- Optional handling of missing values as own group

Data replacement

- Replace nominal / categorical values, also dictionary-based
- Trimming nominal values
- Mapping
- Cutting
- Splitting
- Merging
- Handling missing values by minimum, maximum, average, zero, or a user-specified value
- Imputing missing values by arbitrary modeling methods
- Replacing infinite values
- Fill data gaps

Weighting and selection

- Attribute selection:
 - Remove useless attributes
 - Remove attributes unrelated to target based on a chi-square or correlation-based selection criterion
 - Remove attributes unrelated to target based on arbitrary weighting schemes like information gain, gini index, and many more
 - Remove attributes with many missing values
 - Selecting random subsets
 - Selecting by user specification

- Automatically optimized selections:
 - Evolutionary
 - Forward selection
 - Backward elimination
 - Weight-guided
 - Brute-force
 - LARS and LASSO
- Principal Component Analysis (PCA):
 - Calculates Eigenvalues and Eigenvectors from correlation and covariance matrices
 - Plots for principal components coefficients, Eigenvalues, and cumulative variance of Eigenvalues
 - Choose the number of components to be retained
- Support for independent component analysis (ICA)
- Support for Generalized Hebbian Algorithm (GHA)
- Singular Value Decomposition
- Dimensionality reduction with Self-Organizing Maps (SOM)
- Support for Fast Map
- Correspondence Analysis
- More than 30 additional weighting schemes measuring the influence of attributes and forming the base for weight-based selections (filter approach)

Attribute generation

- Operators for generating IDs, copies, concatenations, aggregations, products, Gaussian distributions, and many more
- Automatically optimized generations and detection of latent variables:
 - Evolutionary weighting
 - Forward weighting

- Backward weighting
- Multiple algorithms for the automatic creation of new attributes based on arbitrary functions of existing attributes
- Genetic programming

Descriptive statistics

- Univariate statistics and plots:
 - Numerical attributes: mean, median, minimum, maximum, standard deviation, and number of missing values
 - Nominal / categorical attributes: number of categories, counts, mode, number of missing values
 - Date attributes: minimum, maximum, number of missing values
 - Distribution plots
- Bivariate statistics and plots:
 - Correlation matrix
 - Covariance matrix
 - Anova matrix
 - Grouped Anova
 - Transition matrix
 - Transition graph
 - Mutual information matrix
 - Rainflow matrix
- Scaled and non-scaled mean-deviation plots
- Plots of attribute weights based on multiple types of connection with targets

Graphs and visualization

- Scatter
- Scatter matrices
- Line
- Bubble
- Parallel
- Deviation
- Box
- Contour

- 3-D
- Density
- Histograms
- Area
- Bar charts
- Stacked bars
- Pie charts
- Survey plots
- Self-organizing maps
- Andrews curves
- Quartile
- Surface plots
- Full support for zooming and panning
- Advanced chart engine for arbitrary definition of multiple charts including on-the-fly grouping, filtering, and aggregation
- Choose from several color schemes
- Simple rescaling of axes
- Charts are fully customizable with colors, titles, footnotes, fonts etc.
- Plots can be easily copied and pasted into other applications or exported in various bitmap and vector-based file formats

Similarities

- Calculation of similarities between data points
- Cross Distances
- Conversion between similarities and data sets and vice versa
- Numerical distance measures:
 - Euclidean
 - Canberra
 - Chebychev
 - Correlation
 - Cosine
 - Dice
 - Dynamic Time Warping
 - Inner product
 - Jaccard

- Kernel-Euclidean
- Manhattan
- Max-Product
- Overlap
- Nominal / categorical distance measures:
 - Nominal
 - Dice
 - Jaccard
 - Kulczynski
 - Rogers-Tanimoto
 - Russel-Rao
 - Simple Matching
- Mixed Euclidean distance for cases with numerical/nominal attributes
- Support for Bregman divergences:
 - Itakura-Saito
 - Kullback-Leibler
 - Logarithmic loss
 - Logistic loss
 - Mahalanobis
 - Squared Euclidean
 - Squared loss

Clustering

- User defined clustering or automatically chooses the best clusters
- Several strategies for encoding class into the clustering
- Supported methods:
 - K-Means for all available distance and similarity measures
 - K-Medoids for all available distance and similarity measures
 - Kernel K-Means
 - X-Means
 - Cobweb
 - Clupe
 - DBScan
 - Expectation Maximization Clustering
 - Support Vector Clustering
 - Self-organizing maps

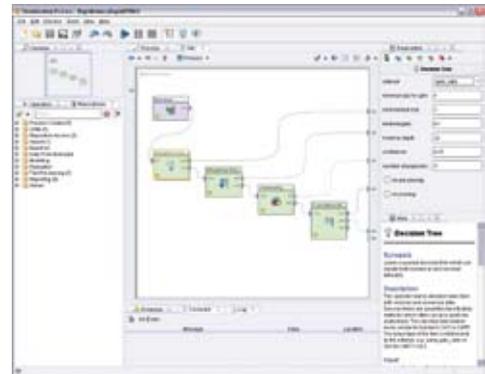
- Agglomerative
- Top Down
- Operators for flatten hierarchical cluster models
- Extraction for prototypes for centroid-based cluster models

Market basket analysis

- Associations and sequence discovery
- Measuring quality of rules by support, confidence, La Place, gain, ps-value, lift, or conviction
- Interactive filter for frequent item sets
- Interactive visualization of association rules as a network graph
- Rules description table
- User defined rule filtering depending on minimum value for the above criteria or matching criteria for specific items
- Apriori
- FP-Growth
- Generalized sequential patterns
- Modular operators for the creation of frequent item sets or association rules only
- Post-processing for the unification of item sets

Decision trees

- Easy-to-understand models
- Supported methods: classification and regression trees (CART), CHAID, decision stumps, ID3, C4.5, Random Forest, bagging and boosting, gradient boosting
- More than 20 different methods for tree creation
- Support for multi-way trees
- Pre-pruning and pruning



- Split criteria include information gain, gain ratio, accuracy, and gini index
- Error-based and confidence-based pruning
- Distribution shown at tree leaves
- Height of distribution bars correlate to number of examples in each leaf
- Majority class shown at tree leaves
- Class counts shown as tool tip at tree leaves
- The darkness of connections correlates with the number of examples on this path
- Graphical and textual representation of trees
- Interactive visualization of trees including selecting and moving of nodes

Rule induction

- Recursive technique with easy-to-read results
- Especially useful for modeling rare events like for subgroup discovery
- Supported methods: rule induc-

tion, single rule induction, single attribute, subgroup discovery, tree to rules, PART, Prism, decision tables, RIPPER

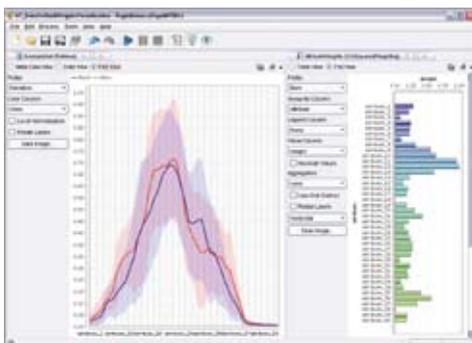
- More than 10 different methods for tree creation
- Supported splitting criteria include information gain and accuracy
- Definition of pureness of rules
- Error-based pruning
- Easy to read and parse representation of rule sets as textual descriptions or tables

Bayesian modeling

- Naïve Bayes
- Kernel Naïve Bayes
- Bayesian Logistic Regression
- Multinomial Naïve Bayes
- Bayes Net
- Complement Naïve Bayes
- Bayes models can be updated and are therefore especially suitable for large data sets or online stream mining

Regression

- Linear
- Logistic
- Kernel Logistic Regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Regularized Discriminant Analysis (RDA)
- Stepwise forward and backward selection



- Selection with M_5' , t-test or iterative t-test
- Seemingly unrelated regression
- Vector linear regression
- Polynomial regression
- Local polynomial regression
- Gaussian Process
- Isotonic
- Partial least squares (PLS)
- Least median squares
- Pace
- RBF network

Neural networks

- Flexible network architectures with different activation functions
- Multiple layers with different numbers of nodes
- Different training techniques
- Perceptron
- Voted Perceptron
- Multilayer Perceptron
- Automatic optimization of both learning rate and size adjustment of neural networks during training

Support vector machines

- Powerful and robust modeling techniques for large numbers of dimensions
- Offers overfitting control by regularization
- Especially suitable modeling unstructured information like text data
- More than 10 different methods for support vector classification, regression, and clustering
- Support vector machine
- Relevance vector machine
- Kernel functions include dot, radial basis function, polynomial, neural, anova, Epachnenikov, Gaussian

- combination, or multiquadric
- Simple support vector machines for boosting support
- Linear-time support vector machine for fast training also for large numbers of dimensions and examples

Memory-based reasoning

- k-nearest neighbors
- Locally weighted learning
- Optimized search in ball trees for faster scorings

Model ensembles

- Hierarchical models
- Combination of multiple models to form a potentially stronger model:
 - Vote
 - Additive regression
 - Ada boost
 - Bayesian boosting
 - Bagging
 - Stacking
 - Decorate
 - Dagging
 - Grading
- Classification by regression
- Meta cost for defining costs for different error types and detecting optimal models avoiding expensive errors

Model evaluation

- Display multiple results in history to help better evaluate model performance
- Various techniques for the estimation of model performance:
 - Cross validation
 - Split validation
 - Bootstrapping

- Batch cross validation
- Wrapper cross validation
- Wrapper split validation
- Visual evaluation techniques:
 - Lift chart
 - ROC curves
 - Confusion matrix
- Many performance criteria for numerical and nominal / categorical targets, including:
 - Accuracy
 - Classification error
 - Kappa
 - Area under curve (AUC)
 - Precision
 - Recall
 - Lift
 - Fallout
 - F-measure
 - False positives
 - False negatives
 - True positives
 - True negatives
 - Sensitivity
 - Specificity
 - Youden index
 - Positive predictive value
 - Negative predictive value
 - PSEP
 - Correlation
 - Spearman rho
 - Kendall tau
 - Squared correlation
 - Absolute error
 - Relative error
 - Normalized absolute error
 - Root mean squared error (RMSE)
 - Root relative squared error (RRSE)
 - Squared error
 - Cross entropy
 - Margin
 - Soft margin loss
 - Logistic loss
- Calculating significance tests to

determine if and which models performed better:

- T-test
- Anova
- Find threshold operator to determine optimal cutoff point for binominal classes
- Performance estimation for cluster models based on distance calculations, density calculations, or item distributions

Scripting

- Write RapidMiner Scripts for easy-to-complex data preparation and transformation tasks where existing operators might not be sufficient
- Incorporate procedures from other processes or projects
- Develop custom models
- Augment scoring logic by custom post-processing or model application procedures
- Easy-to-use program development interface:
 - Predefined imports for common data structures
 - Syntactic sugar for simplified data access and alteration
 - Interactive code editor and syntax highlighting
- Execute arbitrary command line programs and integrate results and result codes in processes
- Execution of SQL statements directly in databases
- Seamless integration of the statistical programming language R into the RapidMiner user interface:
 - Execution of R scripts within RapidMiner processes
 - Predefined R models and transformations available as operators

- Custom R scripts can be stored and executed as own operators directly within a RapidMiner process
- R perspective consisting of the known R console together with the plotting facilities of R
- All R variables as well as R scripts can be stored in the RapidMiner Repository

Process control

- Organize segments in sub-processes and reuse them in different projects
- Repeat execution over a segment of a process
- Support for loops:
 - Attributes
 - Labels
 - Subsets
 - Values
 - Examples
 - Clusters
 - Batches
 - Data Sets
 - Data Fractions
 - Parameters
 - Files
 - Repository entries
- Branches (if-then-else) based on:
 - Data values
 - Attribute existence
 - Numbers of examples
 - Performance values
 - Existence of files and process inputs
 - Definition of macros
 - Arbitrary expressions
- Creation of collections of the same type
- Collection handling: selection, flattening, or looping
- Remembering and recalling (inter-

- mediate) process results for complex process designs
- Handling expected and unexpected errors and exceptions

Automatic optimization

- Automatic selection of best performing sub-processes
- Measuring the influence of preprocessing steps by nested cross validations / other validations
- Automatic selection of best model type and parameters
- Automatic selection of best attribute subsets
- Automatic optimization of process parameters including modeling parameters:
 - Grid
 - Quadratic
 - Evolutionary

Macros

- Centralized definition of macros/ variables containing arbitrary textual or numerical content
- Usage of macros everywhere in the process design, especially as value for parameters
- Macros can be defined during the process or in the process context
- Definition of macros in the context allows for parameterization of complete processes, e.g. for transforming processes into customizable web services
- Extraction of macro values from data values, meta data or statistics supported
- Expression engine for calculating arbitrary macro values from existing macros

Logging

- Logging can be introduced at arbitrary places within a process
- Logging can collect parameter values, performance values, or specific values for each operator, e.g. the current generation for evolutionary algorithms
- Data values can be logged
- Macro values can be logged
- Logged values can be transformed into several formats including data sets and weights which can be stored, transformed, analyzed, or visualized like any other data set

Process-based reporting

- In cases where logging alone is not sufficient, a complete process-based reporting engine allows for the collection of arbitrary results in static reports
- Different formats like PDF, Excel, HTML, or RTF supported
- Different reporting styles including a sequential report or portals
- Arbitrary process results as well as intermediate results can be transformed into different types of visualizations like tables, charts etc.
- Support of sections with up to 5 levels
- Support for page breaks and other style information
- Combination with loops or other process control structures allows for highly-detailed result over views even for complex process designs

Connection to server

- All RapidMiner processes can be executed on the server RapidAnalytics
- True business analytics with data integration, transformation, analysis, data mining, and reporting in a single suite:
 - Allows for more powerful hardware than desktop application alone
 - Remote execution of analysis processes
 - Scheduled execution of analysis processes
 - Powerful user management
 - Shared repository for collaboration of analysts or central storage
 - Cluster support and distributed process execution engine
- Web-based access to reports, results, and processes:
 - Processes can easily be transformed into web services
 - Web services / processes can deliver XML, JSON, or static / dynamic visualizations among others
 - Results can be styled with XSLT
 - Pixel-perfect interactive report designer
 - Interactive dashboards: report elements can be connected with a few clicks only
 - Support for multiple views as part of a single report / dashboard
- Ad-hoc Reporting
- Style bundles
- Customized branding
- Simplified integration through web service executions or iframe-based web integration

- Report elements can be exported for third-party portal systems supporting the JSR-168 standard

Extensions

- Various extensions for RapidMiner and RapidAnalytics exist which add new features or increase productivity
- Big data analytics made easy with Radoop:
 - Radoop is a big data extension for RapidMiner which allows the processing of terabytes and petabytes of data
 - Radoop combines the strengths of RapidMiner with Hadoop, the result is a solution for the graphical creation and execution of workflows for ETL and predictive analytics on Hadoop clusters
 - Hive (distributed data warehouse) and Mahout (distributed analytical algorithms) already integrated
 - User-friendly drag-n-drop interface of RapidMiner builds a powerful and easy-to-use solution for big data analytics
 - Reduced the complexity of big data systems and allows non-technical staff to create analytical big data workflows without custom scripting
- Integration of R:
 - Arbitrary R models and scripts can be seamlessly integrated into RapidMiner processes
 - R perspective offers the known R console together with plotting facilities of R
 - All variables and R scripts can be organized in the RapidMiner repository
- Online help
- Multi-line editing
- Syntax highlighting
- Widely used modeling methods are integrated as operators
- Integration of Weka:
 - Well-known machine learning library Weka completely included
 - More than 100 additional modeling operators
- Text extension:
 - Operators necessary for statistical text analysis
 - Load texts from different data sources or from your data sets including plain texts, HTML, PDF, RTF, and many more
 - Transform texts by a huge set of different filtering techniques including tokenization, stemming, stop word filtering, part of speech tagging, n-grams, dictionaries, and many more
 - Connection to WordNet and other services to clean up texts before processing
- Web extension:
 - Access to internet sources like web pages, RSS feeds, and web services
 - Specific operators for handling the content of web pages
 - Extend structured data with data from the web and combine those data sources for getting new insights and detect chances and risks
- More than 50 extensions available for all types of input formats, data transformations, and analysis including extensions for image mining, audio mining, time series analysis, automatic system creation and many more

Requirements

Client environment

- Microsoft Windows (x86-32): Windows XP, Windows Server 2003, Windows Vista, Windows Server 2008, Windows 7, or
- Microsoft Windows (x64): Windows XP for x64, Windows Server 2003 for x64, Windows Vista for x64, Windows Server 2008 for x64, Windows 7 for x64, or
- Unix 32 or 64 bit systems, or
- Linux 32 or 64 bit systems, or
- Apple Macintosh 32 or 64 bit systems.

RapidMiner and RapidAnalytics need a Java Runtime Environment version 6 (might not be necessary for Windows versions of RapidMiner).

Recommendations

- Since the maximum amount of memory usable by RapidMiner for 32 bit systems is restricted to at most 2 Gb, we generally recommend the usage of a 64 bit systems and operating systems for RapidMiner.
- Although many analysis tasks can be performed with the RapidMiner desktop client already, we generally recommend the usage of the server RapidAnalytics in combination with RapidMiner. Analysis processes are then designed with RapidMiner but executed on the RapidAnalytics server.

Rapid-I GmbH
Stockumer Str. 475
D-44227 Dortmund
Germany

Phone: +49 (0)231 425 786 90
E-Mail: contact@rapid-i.com
www.rapid-i.com

Rapid-I Inc.
15 New England Executive Park
Burlington, MA 01803
USA
Phone: +1 617-401-7708
Fax: +1 617-401-7709
Toll Free: +1 (855) 4 RAPID-I

