A Comparison of Leading Data Mining Tools

John F. Elder IV & Dean W. Abbott
Elder Research

Fourth International Conference
on Knowledge Discovery & Data Mining

Friday, August 28, 1998

New York, New York

Contacting Elder Research

http://www.datamininglab.com

Dr. John F. Elder IV
1006 Wildmere Place
Charlottesville, VA 22901

elder@datamininglab.com
804-973-7673
Fax: 804-995-0064

Dean W. Abbott
3443 Villanova Avenue
San Diego, CA 92122-2310

dean@datamininglab.com
619-450-0313

Tutorial Goals

- Compare and Summarize Data Mining Tools which:
  - Offer multiple modeling and classification algorithms
  - Support project stages surrounding model construction
  - Stand alone
  - Are general-purpose
  - Cost a lot
  - We could get our hands on
- Include some (focused) Desktop Tools

Other Reports: Two Crows, Aberdeen Group, Elder Research (forthcoming), Data Mining Journal

Topics

- Products covered
- Review of algorithms
- Comparative tables of properties
- Screen shots exemplifying qualities
- Summary of distinctives

Caveats

- We don't know *every* tool well (and are sure to have missed some!)
  - Level of exposure noted for each tool
- Our background (biasing our perspective)
  - Very technical, "early adopters"
  - Emphasize solving real-world applications
  - More classification than estimation
- Field of tools is quite dynamic
  - New versions appear regularly

Data Mining Products

Model 1

Tools Evaluated

Product	Company	URL	Version Tested	Our Experience
Clementine	Integral Solutions, Ltd.	http://www.isl.co.uk/clem.html	4	Moderate
Darwin	Thinking Machines, Corp.	http://www.think.com/html/products/products.htm	3.0.1	Moderate
DataCruncher	DataMind	http://www.datamindcorp.com	2.1.1	High
Enterprise Miner	SAS Institute	http://www.sas.com/software/components/miner.html	Beta	Moderate
GainSmarts	Urban Science	http://www.urbanscience.com/main/gainpage.htm	4.0.3	Low
Intelligent Miner	IBM	http://www.software.ibm.com/data/iminer/	2	Low
MineSet	Silicon Graphics, Inc.	http://www.sgi.com/Products/software/MineSet/	2.5	Low
Model 1	Group 1	http://www.unica-usa.com/model1.htm	3.1	Moderate
ModelQuest	AbTech Corp.	http://www.abtech.com	1	Moderate
PRW	Unica Technologies, Inc.	http://www.unica-usa.com/prodinfo.htm	2.5	High
CART	Salford Systems	http://www.salford-systems.com	3.5	Moderate
NeuroShell	Ward Systems Group, Inc.	http://www.wardsystems.com/neuroshe.htm	3	Moderate
OLPARS	PAR Government Systems	mailto://olpars@partech.com	8.1	High
Scenario	Cognos	http://www.cognos.com/busintell/products/index.html	2	Moderate
See5	RuleQuest Research	http://www.rulequest.com/see5-info.html	1.07	Moderate
S-Plus	MathSoft	http://www.mathsoft.com/splus/	4	High
WizWhy	WizSoft	http://www.wizsoft.com/why.html	1.1	Moderate

Categories for Comparisons

- Platforms Supported
- Algorithms Included
  - Decision Trees
  - Neural Networks
  - Other
- Data Input and Model Output Options
- Usability Ratings
- Visualization Capabilities
- Modeling Automation Methods

Platforms	PC Standalone (95/NT)	Unix Standalone	Unix Server / PC Client	NT Server / PC Client	Database Connectivity
Clementine	√	√+			√
Darwin			√		√
DataCruncher	√		√		√
Enterprise Miner	√		√+	√	√
GainSmarts	√	√			√
Intelligent Miner			√		√
MineSet		√			√
Model 1	√			√	√
ModelQuest	√	√			√
PRW	√				√
CART	√	√+			
Scenario	√				√
NeuroShell	√				
OLPARS	√	√			
See5	√	√+			
S-Plus	√				√–
WizWhy	√				

Key

blank	no capability
√–	some capability
√	good capability
√+	excellent capability

Tool Groupings

Desktop

- PC (standalone)
- Flat Files
- One or Two Algorithms
- Data Fits into RAM

High End

- Multiple Platforms, Client-Server
- Flat Files or Direct Database Access
- Multiple Algorithm Types
- Large Databases

End User Perspectives

Business

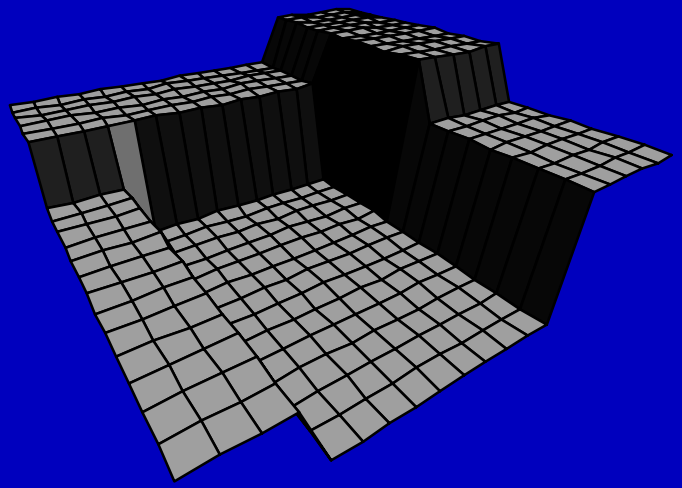
- Intuitive Interface
  - Clear steps in data mining process
  - Non-technical terminology
  - Familiar environment
- Descriptive Reporting
  - Domain terminology
  - Graphical representations

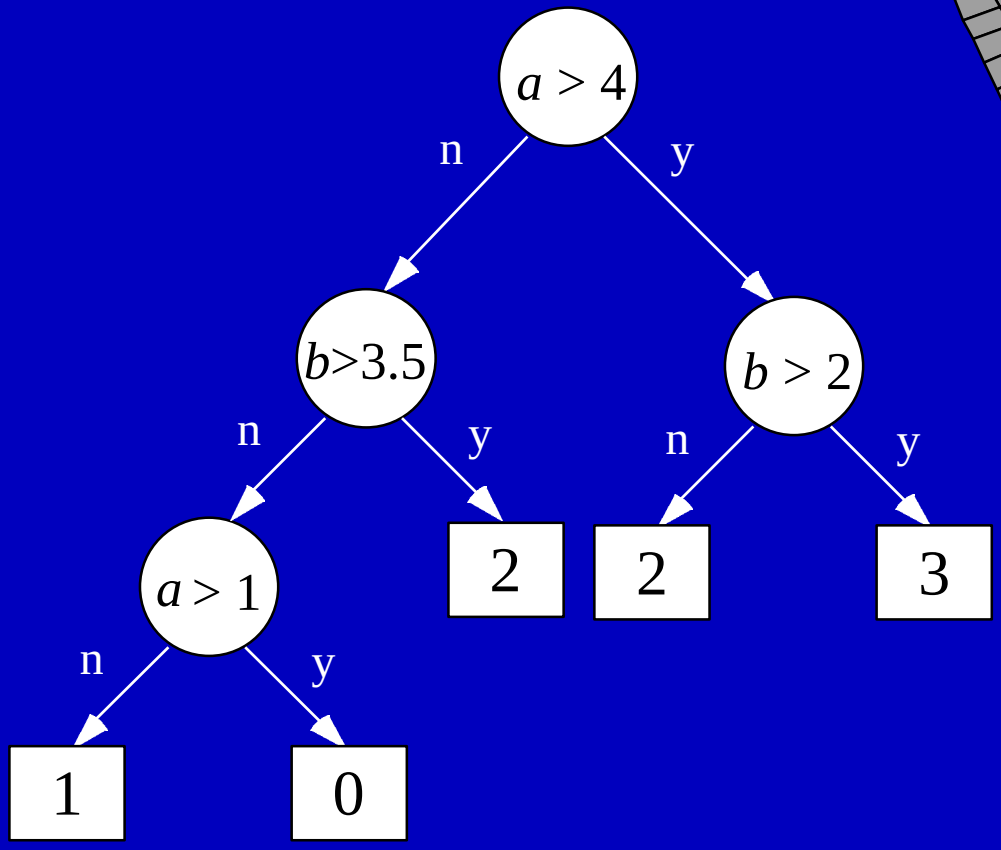
Technical

- Algorithm Options
  - Knobs to enhance model performance
- Model Automation
  - Simplify model design cycle
  - Documentation of steps used in generating models (repeatability)

Data Input & Model Output	Automatic Header	Save Data Format	ODBC	Native Database Drivers	Summary Reports	Output Source Code
Clementine	√		√			√
Darwin		√	√			√
DataCruncher	√	√	√	√	√	
Enterprise Miner	√–		√	√	√–	√
GainSmarts	√	√		√	√	√
Intelligent Miner				√–		√
MineSet		√		√		
Model 1	√	√	√		√	√
ModelQuest	√			√	√	√
PRW	√	√	√			√
CART	√					
Scenario	√				√	
NeuroShell	√					
OLPARS		√				
See5	√–					
S-Plus	√		√		√	√
WizWhy	√				√	

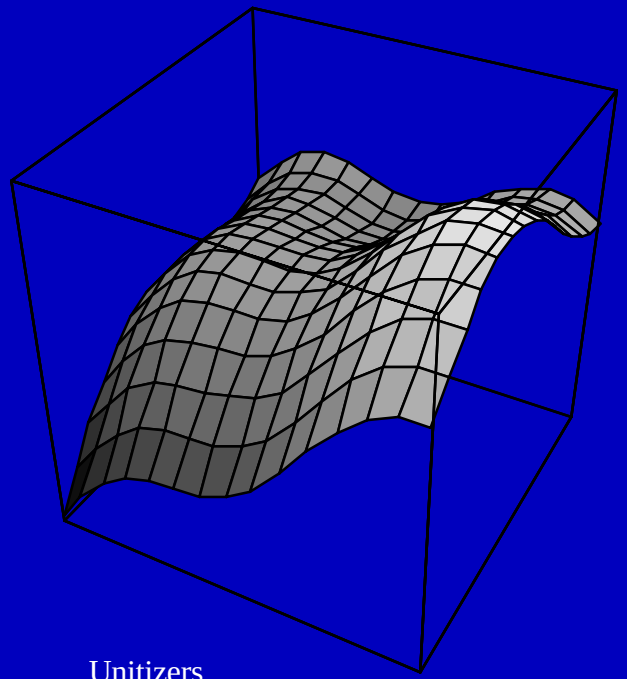
Decision Trees

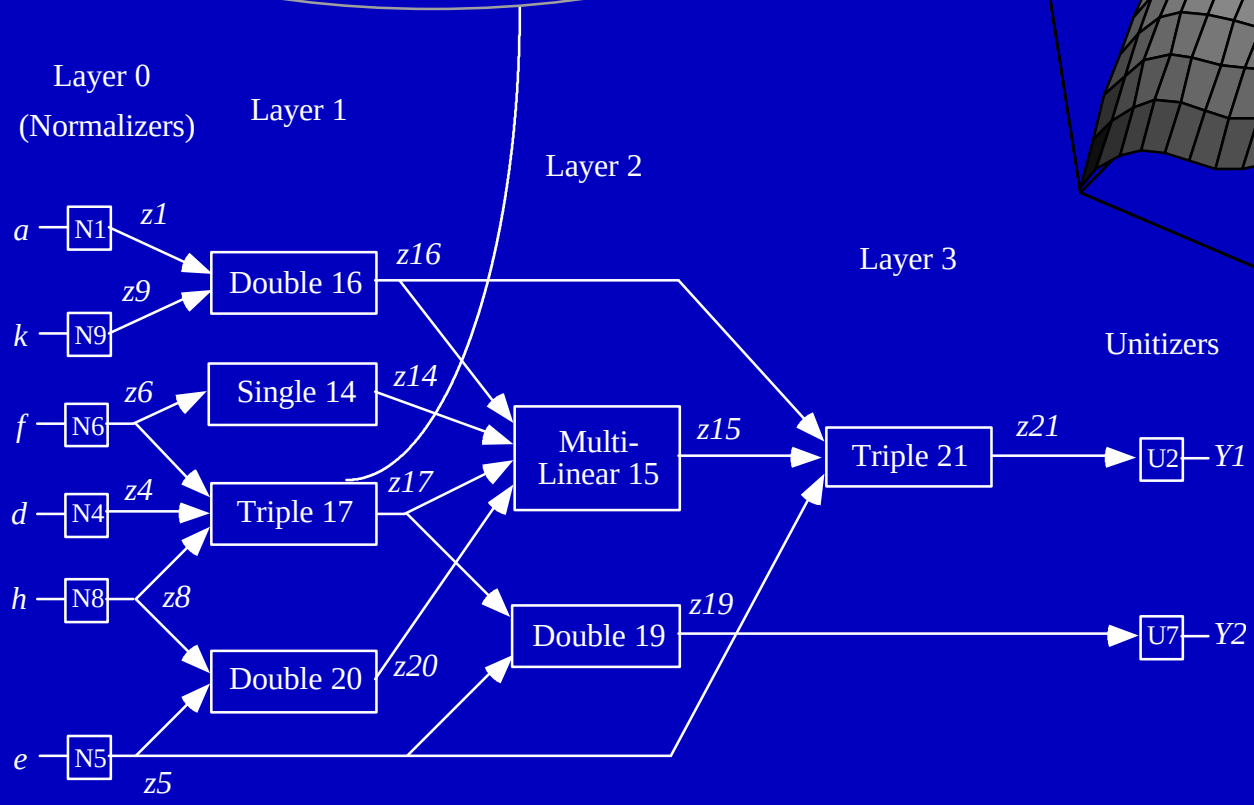




Polynomial Networks

$$Z_{17} = 3.1 + 0.4a - .15b^2 + 0.9bc - 0.62abc + 0.5c^3$$



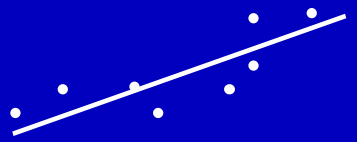


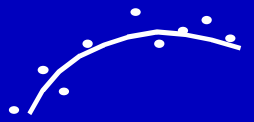
"Consensus" Models

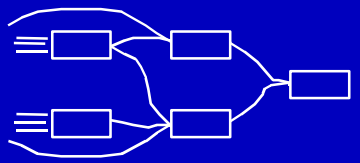
Parametrically Summarize Data Points

orders, terms

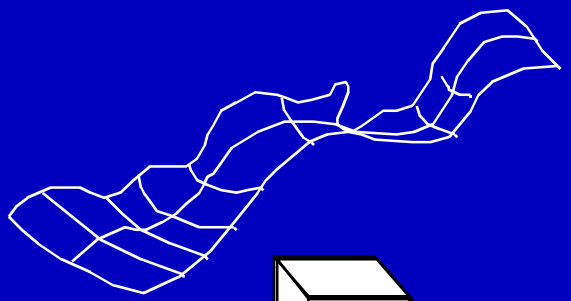
Regression

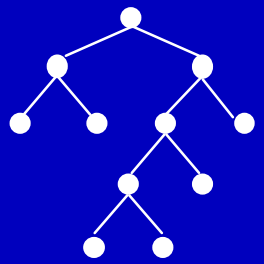




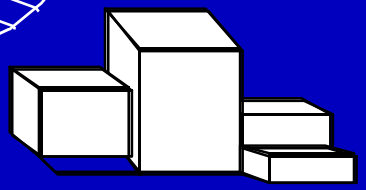


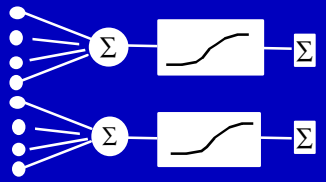
Polynomial Networks
(e.g. GMDH, ASPN)



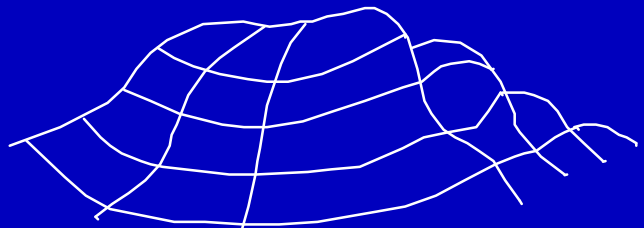


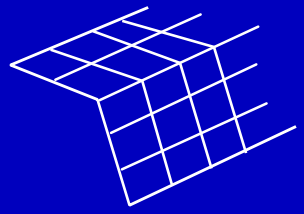
Decision Trees
(e.g., CART, CHAID, C5)



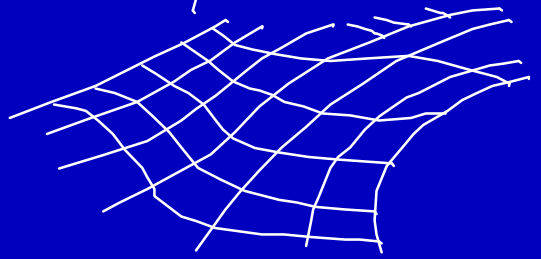


Logistic or Sigmoidal
Networks (ANNs)



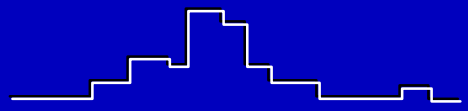


Hinging Hyperplanes,
MARS



"Consensus" Models (continued)

orientation, bin width — Histogram

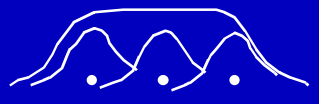
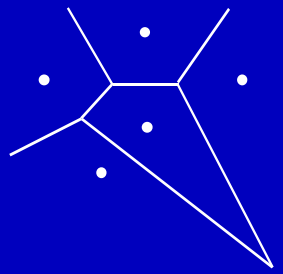
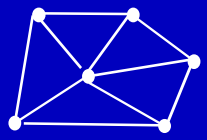
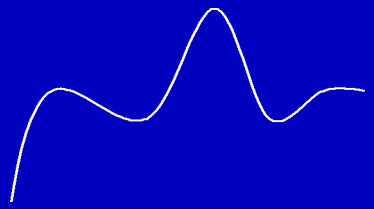


function — Radial Basis Function

family, order — Wavelets

"Contributory" Models

retain data points; each potentially affects estimate at new point

shape, spread	Kernels	
k, distance metric	k-Nearest Neighbor	
Goal, iterations	Delaunay Planes	
Spread, index	Projection Pursuit Regression	

Properties of Algorithms

Algorithm	Accurate	Scalable	Interpret-able	Useable	Robust	Versatile	Fast	Hot
Classical (LR, LDA)	–	👍	👍–	👍	–	–	👍	👎
Neural Networks	👍	👎	👎	👎	–	👎	👎👎	👍
Visualization	👍	👎👎	👍	👍	👍👍	👎	👎👎👎	👍–
Decision Trees	👎	👎	👎	👍–	👎	👎	👍–	👍–
Polynomial Networks	👍	–	👎	👍–	–👎	–	–👎	–
K-Nearest Neighbors	👎	👎👎	👍–	–	–👎	👎	👍	👎
Kernels	👍	👎👎	👎	–👎	👎	👎	👍	👎

Key

👍 good

– neutral

👎 bad

Algorithms	Decision Trees	Linear/Statistical	Multi-layer Perceptrons	Nearest Neighbor	Radial Basis Functions	Bayes	Rule Induction	Polynomial Networks	Generalized Linear Models	Time Series	Sequential Discovery	K Means	Association Rules	Kohonen
Clementine	√	√	√				√					√	√	√
Darwin	√		√	√										
Datamind							√							
Enterprise Miner	√	√	√		√				√	√		√	√	
GainSmarts	√	√+												
Intelligent Miner	√	√–	√		√–					√	√	√+	√	
MineSet	√					√						√	√	
Model 1	√+	√	√											
ModelQuest	√	√		√				√		√–				
PRW		√+	√	√	√	√						√		
CART	√													
Cognos	√													
NeuroShell			√+		√					√–				
OLPARS		√	√	√	√	√						√		√
See5	√						√							
SPlus	√	√+							√	√		√		
WizWhy							√							

Multi-Layer Perceptrons	Learning Rate	Learning Rate Decay	Momentum	Multiple Activation Functions	Multiple Stop Criteria	Cross-Validation	Normalize Inputs	Advanced Learning Alg.	Other Cost functions	Automatic Model Selection	Network Visual	Parameter Summary
Clementine	√	√	√							√		
Darwin	√			√		√		√	√			√
Enterprise Miner	√	√	√	√	√		√	√	√		√	√
Intelligent Miner	√									√		
Model 1	√									√		
PRW	√		√		√		√			√		√
NeuroShell	√	√	√	√	√							
OLPARS	√		√	√	√		√				√	√

Decision Trees	"CART"	C5 or C4.5	CHAID	Other	Priors	Classification Costs	Missing Data	Pruning Severity	Visual Trees
Clementine		√				√	√	√	√–
Darwin	√				√	√	√		
Enterprise Miner	√	√–	√		√+	√	√	√	√
GainSmarts	√		√	√			√		√
Intelligent Miner				√			√		√
MineSet	√		√			√	√	√	√
Model 1	√		√				√–		
ModelQuest		√–					√	√	
CART	√+				√	√	√		√
Scenario				√			√		
S-Plus	√						√	√	√
See5		√+				√	√	√	

Regression / Stats	Linear	Logistic	Complexity Penalty	Cross-Validation	Input Selection	Factor Analysis
Clementine	√					
Enterprise Miner	√+	√+	√	√	√	√
GainSmarts	√+	√+	√			
Intelligent Miner	√–				√	√
MineSet	√					
Model 1	√	√		√	√+	
ModelQuest Enterprise	√	√	√	√	√	
PRW	√	√		√	√+	
S-Plus	√+	√+	√	√	√	√
Scenario						√

Usability	Data Loading and Manipulation	Model Building	Model Understanding	Technical Support	Overall
Clementine	√+	√+	√+	√+	√+
Darwin	√	√	√+	√	√
DataCruncher	√+	√+	√	√	√
Enterprise Miner	√	√	√	√	√
GainSmarts	√+	√	√	√	√
Intelligent Miner	√	√	√	√	√
MineSet	√	√+	√+	√	√+
Model 1	√+	√+	√+	√+	√+
ModelQuest Enterprise	√	√+	√+	√+	√+
PRW	√+	√+	√+	√+	√+
CART	√–	√	√	√	√
Scenario	√	√+	√+	√	√+
NeuroShell	√	√	√	√	√
OLPARS	√–	√	√	√	√
See5	√	√	√	√	√
S-Plus	√	√	√+	√	√
WizWhy	√	√	√+	√	√

Visualization	Histograms	Pie Charts	Scatter/ Line Plots	Rotating Scatter	Conditional Plots	Classification Decision Regions	Correlation Plots
Clementine	√		√		√	√–	√
Darwin	√–	√–	√–				
DataCruncher	√	√	√		√		
Enterprise Miner	√	√	√	√–	√		√
GainSmarts	√–		√–				
Intelligent Miner	√	√	√		√		
MineSet	√	√	√	√	√		
Model 1	√	√	√	√			
ModelQuest Enterpris	√		√				
PRW	√	√	√	√			
CART							
Scenario							√
NeuroShell			√				
OLPARS	√	√	√	√–	√	√	
See5	√						
S-Plus	√	√	√		√		√
WizWhy							

Automation	Method of Automation	Free Text Annotation of Steps
Clementine	Visual Programming, Programming Language	√
Darwin	Programming Language	√
DataCruncher	(Task manager)	
Enterprise Miner	Visual Programming, Programming Language	√
GainSmarts	Macro Language, Wizards	√–
Intelligent Miner	(Wizards)	
MineSet	Data History, Log	
Model 1	Model Wizard	
ModelQuest	Batch Agenda	
PRW	Experiement Manager; Macros	√
CART	Built-in Basic Scripting	
Scenario		
NeuroShell		
OLPARS		
See5		
S-Plus	Scripting (S); C/C++	
WizWhy		

A Recent Breakthrough: Bundling

1) Construct varied models, and
2) Combine their estimates

Generate component models by varying:

- Case Weights
- Data Values
- Guiding Parameters
- Variable Subsets

Combine estimates using:

- Estimator Weights
- Voting
- Advisor Perceptrons
- Partitions of Design Space

Example Bundling Techniques

- *Bayes*: sum estimates of possible models, weighted by priors
- *GMDH* (Ivakhenko 68) -- multiple layers of quadratic polynomials, using two inputs each, fit by LR
- *Stacking* (Wolpert 92) -- train a 2nd-level (LR) model using leave-1-out estimates of 1st-level (neural net) models
- *Bagging* (Breiman 96) (*b*ootstrap *agg*regating) -- bootstrap data (to build trees mostly); take majority vote or average
- *Bumping* (Tibshirani 97) -- bootstrap, select single best
- *Boosting* (Freund & Shapire 96) -- weight error cases by $\beta\tau$ = (1-e(*t*))/e(*t*), iteratively re-model; weight model *t* by ln($\beta\tau$)
- *Crumpling* (Anderson & Elder 98) -- average cross-validations
- *Born-Again* (Breiman 98) -- invent new X data...

Distinctives	Strengths	Weaknesses
Clementine	visual interface; algorithm breadth	scalability
Darwin	efficient client-server; intuitive interface options	no unsupervised; limited visualization
DataCruncher	ease of use	single algorithm
Enterprise Miner	depth of algorithms; visual interface	harder to use; new product issues
GainSmarts	data transformations, built on SAS; algorithm option depth	no unsupervised; limited visualization
Intelligent Miner	algorithm breadth; graphical tree/cluster output	few algorithm options; no automation
MineSet	data visualization	few algorithms; no model export
Model 1	ease of use; automated model discovery	really a vertical tool
ModelQuest	breadth of algorithms	some non-intuitive interface options
PRW	extensive algorithms; automated model selection	limited visualization
CART	depth of tree options	difficult file I/O; limited visualization
Scenario	ease of use	narrow analysis path
NeuroShell	multiple neural network architectures	unorthodox interface; only neural networks
OLPARS	multiple statistical algorithms; class-based visualization	dated interface; difficult file I/O
See5	depth of tree options	limited visualization; few data options
S-Plus	depth of algorithms; visualization; programable/extendable	limited inductive methods; steep learning curve
WizWhy	ease of use; ease of model understanding	limited visualization

Closing Observations

- Data Mining Tools Can:
  - Enhance inference process
  - Speed up design cycle
- Data Mining Tools Can Not:
  - Substitute for statistical and domain expertise
- Users are advised to:
  - Get training on tools
  - Be alert for product upgrades

Forthcoming Report

- Report provides detailed comparison of high-end data mining tools, including capabilities, ease of use, and practical tips.
- Available for $695 from Elder Research (http://www.datamininglab.com), Q4 1998.
- Purchasers receive brief free consulting session to explore report findings in more detail, if desired.

Note: The analyses and reviews were performed completely independently, and were made possible by the cooperation of the vendors, for which Elder Research is very grateful. The companies, however, provided no financial support, and had no influence on its editorial content.