Data Mining and Computation Software for Improving Fisheries Research

Geethalakshmi V. & Chandrasekar V.
Extension, Information and Statistics Division
ICAR-Central Institute of Fisheries Technology, Cochin

Introduction

The branch of science which deasls with data generation, management, analysis and information retrieval is called "Statistics". Statistics methods and advanced computational techniques are very important and crucial to fisheries research and management. Statistics has a key role to play in fisheries research carried out in the various disciplines viz., Aquaculture, Fisheries Resource Management, Fish Genetics, Fish Biotechnology, Aquatic Health, Nutrition, Environment, Fish Physiology and Post-Harvest Technology for enhancing production and ensuring sustainability. For formulating advisories and policies for stakeholders at all levels, the data generated from the various sub-sectors in fisheries and aquaculture has to be studied.

Statistical system can play more dominant role

- in providing tools for policy making and implementation
- in directing the impact of technology
- in sustaining the nutritional safety
- in socio-economic upliftment of people below poverty line
- to identify emerging opportunities through effective coordination
- speedy dissemination of information by networking and appropriate human resource development

Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. The key properties of data mining are

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information
- Focus on large datasets and databases

Data mining can generate new business opportunities when databases of sufficient size and quality, are analysed for patterns, and based on this action can be streamlined. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return

on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

Data mining is at the heart of analytics efforts across a variety of industries and disciplines. Telecom, media and technology In an overloaded market where competition is tight, the answers are often within your consumer data. Telecom, media and technology companies can use analytic models to make sense of mountains of customers data, helping them predict customer behaviour and offer highly targeted and relevant campaigns.

Education With unified, data-driven views of student progress, educators can predict student performance before they set foot in the classroom – and develop intervention strategies to keep them on course. Data mining helps educators access student data, predict achievement levels and pinpoint students or groups of students in need of extra attention.

Finance & banking Automated algorithms help banks understand their customer base as well as the billions of transactions at the heart of the financial system. Data mining helps financial services companies get a better view of market risks, detect fraud faster, manage regulatory compliance obligations and get optimal returns on their marketing investments.

Insurance With analytic know-how, insurance companies can solve complex problems concerning fraud, compliance, risk management and customer attrition. Companies have used data mining techniques to price products more effectively across business lines and find new ways to offer competitive products to their existing customer base.

Manufacturing Aligning supply plans with demand forecasts is essential, as is early detection of problems, quality assurance and investment in brand equity. Manufacturers can predict wear of production assets and anticipate maintenance, which can maximize uptime and keep the production line on schedule.

Retailing Large customer databases hold hidden customer insight that can help you improve relationships, optimize marketing campaigns and forecast sales. Through more accurate data models, retail companies can offer more targeted campaigns – and find the offer that makes the biggest impact on the customer.

Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Data can be of the following types - Record data – Transactional, Temporal data – Time series, sequence (biological sequence data), Spatial & Spatial-Temporal data, Graph data, Unstructured data -twitter, status, review, news article and Semi-structured data -publication data, xml. Data mining can be employed for:

Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.

Association rule learning (Dependency modelling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.

Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".

Regression – attempts to find a function which models the data with the least error.

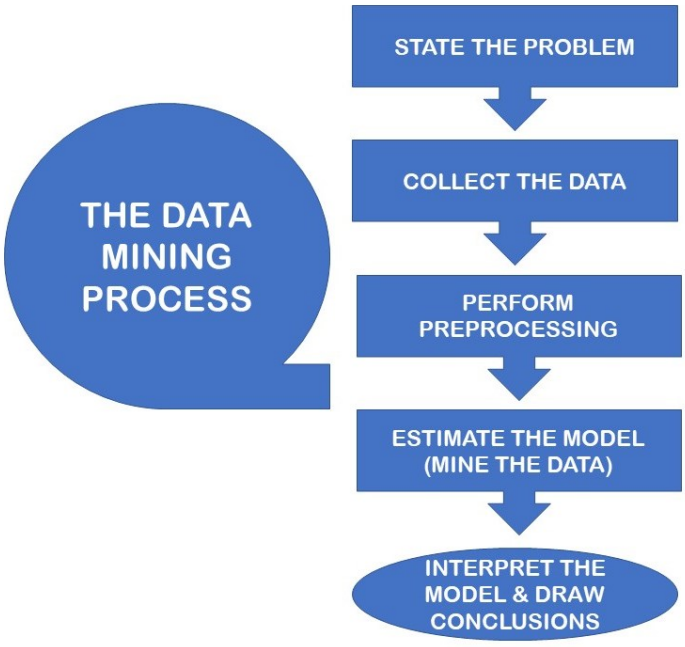
Summarization – providing a more compact representation of the data set, including visualization and report generation.

The Data Mining Process

In order to explore the unknown underlying dependency in the data an initial hypothesis is assumed. There may be several hypotheses formulated for a single problem at this stage. Data generation is the second step which can be either through a designed experiment. The second possibility is when the expert cannot influence the data- generation process: this is known as the observational approach. An observational setting, namely, random data generation, is assumed in most data-mining applications. Data collection affects its theoretical distribution. It is important to make sure that the data used for estimating a model and the data used later for testing and applying a model come from the same, unknown, sampling distribution. In the observational setting, data are usually "collected" from the existing databases, data warehouses, and data marts.

Data pre-processing is an important step before doing the analysis. Firstly outliers have to be identified and removed or treated. Commonly, outliers result from measurement errors, coding and recording errors, and, sometimes, are natural, abnormal values. Such nonrepresentative samples can seriously affect the model produced later. Pre-processing involves either removal of outliers from data or develop robust models which are insensitive to outliers. Data pre-processing also includes several steps such as variable scaling and different types of encoding. For estimating the model, selection and implementation of the appropriate data-mining technique is an important step.

Data-mining models should help in decision making. Hence, such models need to be interpretable in order to be useful because humans are not likely to base their decisions on complex "black-box" models. Note that the goals of accuracy of the model and accuracy of its interpretation are somewhat contradictory.



Usually, simple models are more interpretable, but they are also less accurate. Modern data-mining methods are expected to yield highly accurate results using high dimensional models.

Data Mining Techniques

Important data mining techniques are

- Classification analysis. This analysis is used to retrieve important and relevant information about data, and metadata
- Association rule learning
- Anomaly or outlier detection
- Clustering analysis
- Regression analysis

Association analysis is the finding of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for a market basket or transaction data analysis. Association rule mining is a significant and exceptionally dynamic area of data mining research. One method of association-based classification, called associative classification, consists of two steps. In the main step, association instructions are generated using a modified version of the standard association rule mining algorithm known as A priori. The second step constructs a classifier based on the association rules discovered.

Classification is the processing of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. Data Mining has a different type of classifier:

- Decision Tree - a flow-chart-like tree structure, where each node represents a test on an attribute value, each branch denotes an outcome of a test, and tree leaves represent classes or class distributions.
- SVM(Support Vector Machine) - is a supervised learning strategy used for classification and additionally used for regression. When the output of the support vector machine is a continuous value, the learning methodology is claimed to perform regression; and once the learning methodology will predict a category label of the input object, it's known as classification.
- Generalized Linear Models - is a statistical technique, for linear modeling.GLM provides extensive coefficient statistics and model statistics, as well as row diagnostics. It also supports confidence bounds.
- Bayesian classification - is a statistical classifier. They can predict class membership probabilities, for instance, the probability that a given sample belongs to a particular class. Bayesian classification is created on the Bayes theorem.
- Classification by Backpropagation
- K-NN Classifier - The k-nearest neighbor (K-NN) classifier is taken into account as an example-based classifier, which means that the training documents are used for comparison instead of an exact class illustration, like the class profiles utilized by other classifiers.
- Rule-Based Classification - represent the knowledge in the form of If-Then rules. An assessment of a rule evaluated according to the accuracy and coverage of the classifier. If more than one rule is triggered then we need to conflict resolution in rule-based classification.
- Frequent-Pattern Based Classification - (or FP discovery, FP mining, or Frequent itemset mining) is part of data mining. It describes the task of finding the most frequent and relevant patterns in large datasets.
- Rough set theory - can be used for classification to discover structural relationships within imprecise or noisy data. It applies to discrete-valued features. Continuous-valued attributes must therefore be discrete prior to their use. Rough set theory is based on the establishment of equivalence classes within the given training data.
- Fuzzy Logic - Rule-based systems for classification have the disadvantage that they involve sharp cut-offs for continuous attributes. Fuzzy Logic is valuable for data mining frameworks performing grouping /classification. It provides the benefit of working at a high level of abstraction.

Clustering Unlike classification and prediction, which analyze class-labelled data objects or attributes, clustering analyzes data objects without consulting an identified class label. In general, the class labels do not exist in the training data simply because they are not known to begin with. Clustering can be used to generate these labels. The objects are clustered based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. That is, clusters of objects are created so that objects inside a cluster have high similarity in contrast with each other, but are different objects in other clusters. Each Cluster

that is generated can be seen as a class of objects, from which rules can be inferred. Clustering can also facilitate classification formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

Regression can be defined as a statistical modelling method in which previously obtained data is used to predicting a continuous quantity for new observations. This classifier is also known as the Continuous Value Classifier. There are two types of regression models: Linear regression and multiple linear regression models.

Data Generation in Fisheries

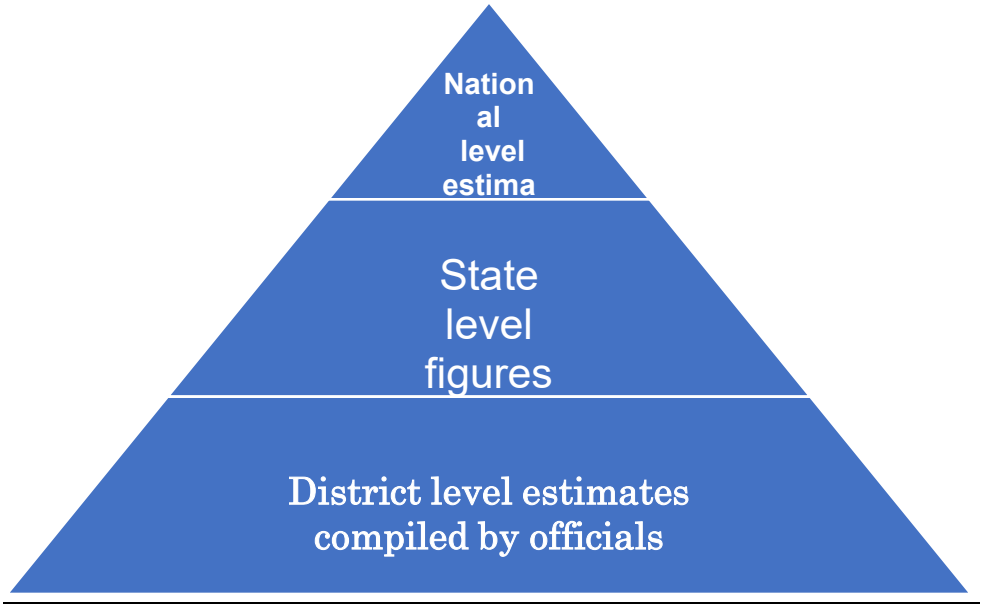
Date generation in fisheries will vary depending on the nature of research undertaken. For example, when species behaviour, growth, abundance, etc. is studied detailed data on spatial distribution and catch is required. If the focus is to predict the profit of the coming years, an economist should study the effect of population size on producer's costs. The macro level data on infrastructure, employment, earnings, investment etc. will be considered to formulate management measures. Enormous data from marine fishing gets generated from commercial fishing vessels and research vessels which will can be mined to analyse the trend, resource abundance, etc.

In 'Fishery technology' large volumes of data generated in a wide range of applied scientific areas of fishing technology, fish processing, quality control, fishery economics, marketing and management. Apart from statistical data collected in technological research, data also collected on production, export, socio-economics etc. for administrative and management decision making.

Major areas of data generation:

- fishing vessel and gear designs
- fishing methods
- craft and gear materials
- craft and gear preservation methods
- fishing efficiency studies
- fishing accessories
- emerging areas include use of GIS and remote sensing

Data on various aspects of fishing gets collected for administrative purposes and policy making. For administrative purposes, voluminous data gets generated through fisheries departments of states. Each district has officials entrusted with the work of collection of data which are coordinated at the state level. State level figures are compiled at the National level by Department of Animal Husbandry and Dairying, Ministry of Agriculture, New Delhi.



Information is also compiled on macro economic variables like GSDP from fishing by the respective Directorates of Economics & Statistics.

Infrastructure

Indian fisheries is supported by a vast fishing fleet of 2,03,202 fishing crafts categorized into mechanized, motorized and non-motorised. The registration of these fishing crafts are done at various ports across India and license for fishing operations has to be obtained from the respective states. The fish processing sector largely managed by the private sector has per day processing capacity installed at 11000 tonnes per day. Data is also collected on the infrastructure facilities and inventories by agencies from time to time such as number of mechanized, motorized and non-motorized fishing crafts, fish landing centers, fisheries harbours, types of gears and accessories, fish markets, ice plants and cold storages, Socio-economic data like population of fishermen, welfare schemes, cooperative societies, financial assistance, subsidies, training programs, etc.

Fish Landings and fishing effort

Indian fisheries has seen tremendous development over the past six decades owing to technology changes in fishing like mechanization of propulsion, gear and handling, introduction of synthetic gear materials, development of acoustic fish finding devices, satellite based fish detection techniques, advances in electronic navigation and communication equipment. The increase in fish production can be said as exponential with a mere 75000 MT in 1950-51 to 11.42 million MT in the current year. Both marine fisheries and aquaculture have contributed to the present level of production with share from culture fisheries more than the capture fisheries. It is important task to collect macro level data from state and country on fish production and details of the species caught in the sea.

The data on fish catch and effort (a measure of fishing activity of vessels at sea), from all the coastal states, Union territories, Islands is being done by ICAR-Central Marine Fisheries

Research institute and maintained as database. Based on standard sampling methodology developed by CMFRI, daily data on commercial landings from selected centres/zones all over the coast is collected, compiled and published. Detailed time series data has been generated on species wise, region wise, gear wise fish landings are collected and compiled for the use of researchers and policy makers. The beach price of fish (species wise) is also collected periodically.

Data on fish farms, production and area under aquaculture is maintained by the respective State Fisheries departments and compiled at the National level. Apart from capture fisheries (marine) and culture fisheries (aquaculture) the fish production from inland water bodies like lake, ponds, reservoirs, etc. is collected and compiled at State level. For developing the sector, various programmes and projects have to be formulated and implemented. To achieve the objectives of such developmental programmes, the current status of production of fish from various regions has to be made known. The need for fish production data maintained by these agencies from marine sources, aquaculture and inland water bodies arises while formulating various research studies and development projects at district, state and National level.

Data generation along the fish value chain

Fresh fish after harvest is iced and distributed through various channels into the domestic markets and overseas markets. Around 80% of the fish is marketed fresh, 12% of fish gets processed for the export sector, 5% is sent for drying/curing and the rest is utilized for other purposes.

Marine Products Export Development Authority (MPEDA) maintains the database on export of fish and fishery products from India to various country. The weekly prices realized by Indian seafood products in the various overseas markets are also collected and compiled by the agency. Marine Products Export Development Authority (MPEDA) established in 1972 under the Ministry of Commerce responsible for collecting data regarding production and exports, apart from formulating and implementing export promotion strategies. Prior to the establishment of MPEDA, Export Promotion Council of India was undertaking this task.

Fish processing factories established all over the country generate data on daily production, procurement of raw material and movement of price structure etc. which is generally kept confidential.Data on quality aspects maintained by Export Inspection Council of India through Export Inspection Agency (EIA) in each region, under Ministry of Commerce and Industry. The EIA is the agency approving the suitability of the products for export.

- bacteriological organisms present in the products
- rejections in terms of quantity
- reason for rejection etc.

Fish quality control

Other types of data generated by CIFT in fishing and fish processing technology are quality control data on fish and fishery products, ice, water, etc. Offshoot of processing technology is Quality Control of which Statistical Quality Control forms an integral part. Due to the stringent quality control measures imposed by importing countries, especially the EU and USFDA standards samples of fish and related products like raw materials, ice and water samples and swabs from fish processing factories are tested at the quality control labs. Another area where statistics gets generated is in product development : consumer acceptability and preference studies mainly for value-added products. Using statistical sensory evaluation methods this data gets analysed.

At Central Institute of Fisheries Technology (CIFT) we are periodically collecting data on the following aspects which is used for policy decisions

- Techno-economic data on various technologies developed
- Data on Economics of operation of mechanized, motorized and traditional crafts
- Data for the estimation of fuel utilization by the fishing industry
- Year wise data on Installed capacity utilization in the Indian seafood processing industry
- Demand – supply and forecast studies on the fishing webs
- Harvest and post-harvest losses in fisheries
- Transportation of fresh fish and utilization of trash fish
- Impact of major trade policies like impact of anti-dumping, trend analysis of price movement of marine products in the export markets
- Study on impact of technology and study on socio-economic aspects

Computational Software for Fisheries Research

R is an open source software that provides a programming environment for doing statistical data analysis. R can be effectively used for data storage, data analysis and a variety of graphing functions. R works on the principle of 'functions' and objects. There are about 25 packages supplied with R (called "standard" and "recommended" packages) and many more are available through the CRAN family of Internet sites (via https://CRAN.R-project.org) and elsewhere. It gets widely used for analysing fisheries data. Also SAS software which

Data Mining Software

Compared to other data mining software, SAS Enterprise Miner is a very comprehensive tool that can handle a wide variety of data mining tasks. Further, it is very user-friendly and easy to learn, even for users who are not familiar with SAS programming. Finally, it has a wide range of built-in features and functionality, which makes it a very powerful tool.

Data mining using SAS

SAS Enterprise Miner is a software tool from SAS that is used for data mining and predictive modelling. It provides a graphical user interface for easy access to a variety of data mining

and machine learning algorithms, and can be used to build predictive models from data sets of any size.

Features:

- SAS Data mining tools help you to analyze Big data
- It is an ideal tool for Data mining, text mining & optimization.
- SAS offers distributed memory processing architecture which is highly scalable

The process flow of SAS Enterprise Miner is as follows:

1. Data is imported into the project.
2. A model is created using the data.
3. The model is validated and deployed.

Data preparation

Data input

You can load a dataset into SAS Enterprise Miner by using the Data Import node. This node allows you to specify the location of the dataset, as well as any other necessary information such as variable types and roles. Nodes are the building blocks of a SAS Enterprise Miner process flow. There are a variety of node types, each of which performs a different task. For example, there are nodes for data import, data cleansing, modelling, and results visualization. The main components of SAS Enterprise Miner are the data source, the data target, the model, and the results. The data source is the location from which the data is being imported. The data target is the location to which the data is being exported. The model is the statistical or machine learning model that is being used to analyze the data. The results are the output of the model, which can be used to make predictions or decisions.

Decision trees are a type of predictive modeling that can be used to classify data. In SAS Enterprise Miner, decision trees are generated using the Tree Model node. This node takes a dataset as input and generates a decision tree based on the variables in the dataset. The tree can then be used to predict the class of new data.

Data partition

You can split datasets in SAS Enterprise Miner by using the Partition node. This node will take a dataset as input and will output two or more partitions, based on the settings that you specify. You can specify the percentage of records that should go into each partition, or you can specify a particular variable to split the dataset on. Partitioning provides mutually exclusive data sets. Two or more mutually exclusive data sets share no observations with each other. Partitioning the input data reduces the computation time of preliminary modelling runs.

The Data Partition node enables you to partition data sets into training, test, and validation data sets. The training data set is used for preliminary model fitting. The validation data set is used to monitor and tune the model weights during estimation and is also used for model assessment. The test data set is an additional hold-out data set that you can use for model

assessment. This node uses simple random sampling, stratified random sampling, or user-defined partitions to create partitioned data sets.

Filtering data

The Filter node tool is located on the Sample tab of the Enterprise Miner tools bar. Use the Filter node to create and apply filters to your training data set. You can also use the Filter node to create and apply filters to the validation and test data sets. You can use filters to exclude certain observations, such as extreme outliers and errant data that you do not want to include in your mining analysis. Filtering extreme values from the training data tends to produce better models because the parameter estimates are more stable.

Explore Node of SAS Enterprise miner

Association node enables you to identify association relationships within the data. For example, if a customer buys a loaf of bread, how likely is the customer to also buy a gallon of milk? The node also enables you to perform sequence discovery if a sequence variable is present in the data set. • The Cluster node enables you to segment your data by grouping observations that are statistically similar. Observations that are similar tend to be in the same cluster, and observations that are different tend to be in different clusters. The cluster identifier for each observation can be passed to other tools for use as an input, ID, or target variable. It can also be used as a group variable that enables automatic construction of separate models for each group.

DMDB node creates a data mining database that provides summary statistics and factor-level information for class and interval variables in the imported data set. The DMDB is a metadata catalog used to store valuable counts and statistics for model building.

Graph Explore node is an advanced visualization tool that enables you to explore large volumes of data graphically to uncover patterns and trends and to reveal extreme values in the database. For example, you can analyze univariate distributions, investigate multivariate distributions, and create scatter and box plots and constellation and 3-D charts. Graph Explore plots are fully interactive and are dynamically linked to highlight data selections in multiple views.

Link Analysis node transforms unstructured transactional or relational data into a model that can be graphed. Such models can be used to discover fraud detection, criminal network conspiracies, telephone traffic patterns, website structure and usage, database visualization, and social network analysis. Also, the node can be used to recommend new products to existing customers.

Market Basket node performs association rule mining over transaction data in conjunction with item taxonomy. This node is useful in retail marketing scenarios that involve tens of thousands of distinct items, where the items are grouped into subcategories, categories,

departments, and so on. This is called item taxonomy. The Market Basket node uses the taxonomy data and generates rules at multiple levels in the taxonomy.

MultiPlot node is a visualization tool that enables you to explore larger volumes of data graphically. The MultPlot node automatically creates bar charts and scatter plots for the input and target variables without making several menu or window item selections. The code created by this node can be used to create graphs in a batch environment.

Path Analysis node enables you to analyze Web log data to determine the paths that visitors take as they navigate through a website. You can also use the node to perform sequence analysis.

SOM/Kohonen node enables you to perform unsupervised learning by using Kohonen vector quantization (VQ), Kohonen self-organizing maps (SOMs), or batch SOMs with Nadaraya-Watson or local-linear smoothing. Kohonen VQ is a clustering method, whereas SOMs are primarily dimension-reduction methods.

StatExplore node is a multipurpose node that you use to examine variable distributions and statistics in your data sets. Use the StatExplore node to compute standard univariate statistics, to compute standard bivariate statistics by class target and class segment, and to compute correlation statistics for interval variables by interval input and target. You can also use the StatExplore node to reject variables based on target correlation.

Variable Clustering node is a useful tool for selecting variables or cluster components for analysis. Variable clustering removes collinearity, decreases variable redundancy, and helps reveal the underlying structure of the input variables in a data set. Large numbers of variables can complicate the task of determining the relationships that might exist between the independent variables and the target variable in a model. Models that are built with too many redundant variables can destabilize parameter estimates, confound variable interpretation, and increase the computing time that is required to run the model. Variable clustering can reduce the number of variables that are required to build reliable predictive or segmentation models.

Variable Selection node enables you to evaluate the importance of input variables in predicting or classifying the target variable. The node uses either an R2 or a Chi-square selection (tree based) criterion. The R-square criterion removes variables that have large percentages of missing values, and remove class variables that are based on the number of unique values. The variables that are not related to the target are set to a status of rejected. Although rejected variables are passed to subsequent tools in the process flow diagram, these variables are not used as model inputs by modelling nodes such as the Neural Network and Decision Tree tools.

Modelling Data using SAS Enterprise miner

AutoNeural node can be used to automatically configure a neural network. The AutoNeural node implements a search algorithm to incrementally select activation functions for a variety of multilayer networks.

Decision Tree node enables you to fit decision tree models to your data. The implementation includes features found in a variety of popular decision tree algorithms (for example, CHAID, CART, and C4.5). The node supports both automatic and interactive training. When you run the Decision Tree node in automatic mode, it automatically ranks the input variables based on the strength of their contribution to the tree. This ranking can be used to select variables for use in subsequent modelling. You can override any automatic step with the option to define a splitting rule and prune explicit tools or subtrees. Interactive training enables you to explore and evaluate data splits as you develop them.

Dmine Regression node enables you to compute a forward stepwise least squares regression model. In each step, the independent variable that contributes maximally to the model R-square value is selected. The tool can also automatically bin continuous terms.

DMNeural node is another modelling node that you can use to fit an additive nonlinear model. The additive nonlinear model uses bucketed principal components as inputs to predict a binary or an interval target variable with automatic selection of an activation function.
Ensemble node enables you to create new models by combining the posterior probabilities (for class targets) or the predicted values (for interval targets) from multiple predecessor models.

Gradient Boosting node uses tree boosting to create a series of decision trees that together form a single predictive model. Each tree in the series is fit to the residual of the prediction from the earlier trees in the series. The residual is defined in terms of the derivative of a loss function. For squared error loss with an interval target, the residual is simply the target value minus the predicted value. Boosting is defined for binary, nominal, and interval targets.
LARS node enables you to use Least Angle Regression algorithms to perform variable selection and model fitting tasks. The LARs node can produce models that range from simple intercept models to complex multivariate models that have many 65 inputs. When using the LARs node to perform model fitting, the node uses criteria from either least angle regression or the LASSO regression to choose the optimal model.

MBR (Memory-Based Reasoning) node enables you to identify similar cases and to apply information that is obtained from these cases to a new record. The MBR node uses k-nearest neighbor algorithms to categorize or predict observations.

Model Import node enables you to import models into the SAS Enterprise Miner environment that were not created by SAS Enterprise Miner. Models that were created by using SAS PROC LOGISTIC (for example) can now be run, assessed, and modified in SAS Enterprise Miner.

Neural Network node enables you to construct, train, and validate multilayer feedforward neural networks. Users can select from several predefined architectures or manually select input, hidden, and target layer functions and options.

Partial Least Squares node is a tool for modelling continuous and binary targets based on SAS/STAT PROC PLS. The Partial Least Squares node produces DATA step score code and standard predictive model assessment results.

Regression node enables you to fit both linear and logistic regression models to your data. You can use continuous, ordinal, and binary target variables. You can use both continuous and discrete variables as inputs. The node supports the stepwise, forward, and backward selection methods. A point-and-click interaction builder enables you to create higher-order modelling terms.

Rule Induction node enables you to improve the classification of rare events in your modelling data. The Rule Induction node creates a Rule Induction model that uses split techniques to remove the largest pure split node from the data. Rule Induction also creates binary models for each level of a target variable and ranks the levels from the most rare event to the most common. After all levels of the target variable are modelled, the score code is combined into a SAS DATA step.

Two Stage node enables you to compute a two-stage model for predicting a class and an interval target variable at the same time. The interval target variable is usually a value that is associated with a level of the class target.

Survival data mining

Survival data mining is the application of survival analysis to data mining problems that concern customers.. The application to the business problem changes the nature of the statistical techniques. The issue in survival data mining is not whether an event will occur in a certain time interval, but when the next event will occur. The SAS Enterprise Miner Survival node is located on the Applications tab of the SAS Enterprise Miner tool bar. The Survival node performs survival analysis on mining customer databases when there are time-dependent outcomes. The time-dependent outcomes are modelled using multinomial logistic regression. The discrete event time and competing risks control the occurrence of the time-dependent outcomes.

The Survival node includes functional modules that prepare data for mining, that expand data to one record per time unit, and perform sampling to reduce the size of the expanded data without information loss. The Survival node also performs survival model training, validation, scoring, and reporting.