SOLUTION BRIEF

NXG Logic Solutions for Chemoinformatics

Solutions

Data Pre-processing

Select from a variety of transformations, dimension reduction methods, fast wavelet transforms, and superresolution ROOT MUSIC analysis.

Machine Learning

Perform knowledge discovery, class discovery, and class prediction, based on machine learning and computational intelligence approaches. Text mining can also be performed for document and concept clustering.

Simulation and Monte Carlo Analysis

NXG Logic technologies provide data fitting and simulation capabilities for more than 20 probability distributions, and Monte Carlo uncertainty analysis.

NXG Logic Advantages

- Time savings
- Transpose-free results
- Manifold learning, knowledge discovery
- Machine learning
- Upgradability

Customers

- Students/interns
- Researchers
- Clinicians
- Faculty
- Engineers
- Data analysts
- Healthcare economists
- Risk managers

Research Fields

- Applied computer science
- Chemistry/chemoinformatics
- Molecular biology/genomics
- Drug design/manufacturing
- Clinical trials
- Bioinformatics
- Medical informatics
- Quality assurance & health outcomes

Chemoinformatics is an information retrieval technique which attempts to evaluate, among other things, quantitative structural activity relationships (QSAR) between molecules of varying structure using secondary sources of information that are supplemental to molecular structure. SMILES strings, or the Simplified Molecular-Input Line-Entry System, is a standard chemical nomenclature for representing molecular structure with ASCII text strings. In its simplest form, text mining usually involves K-means cluster analysis of documents to agglomerate together documents having similar term frequencies. Understanding the cluster structure of documents can reveal the major groups of documents present and the concepts portrayed by each group.

The simplest text mining approach involves membership prediction for a particular document cluster using input text for a single document, and then using this information in predictive analytics derived from data mining. For example, text data from customer complaints logged at call centers can be mined for predicting customers who may churn (go away) or return, followed by distribution of custom-tailored advertisement for increased profitability. Hospitals can use text data from customer satisfaction questionnaires to identify inefficiencies in patient care, or identify correlations between worker satisfaction and customer care in specific work units. NXG Logic technology enables you to perform these text mining analyses.

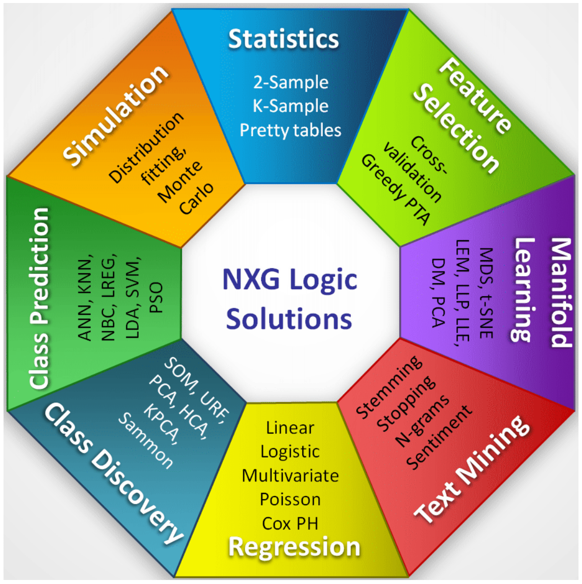
Time-savings. NXG Logic solutions are designed from the bottom-up to shorten and accelerate the time to discovery. Results from each step are typically saved, eliminating the time required for clamping data to the next algorithm. Many graphs are also automatically generated in order to accelerate interpretation.

Transpose-free results. Stop wasting time transposing results from summary statistics and numerous hypothesis tests into user-reader-audience-friendly tables for dissemination, publication, or presentation. NXG Logic summary statistics and hypothesis testing algorithms automatically evaluate numerous tests of assumptions and determine the appropriate tests to be applied. Relevant graphics are also automatically generated.

Manifold learning and knowledge discovery. Most analysts directly input data into hypothesis testing, without sufficiently analyzing the data to determine whether certain patterns exist independent of the experimental group or treatment assignments. NXG Logic users can rapidly transform data, reduce dimensions, perform knowledge and class discovery with cluster validity analysis to identify whether a rich cluster structure of the data exists.

Machine learning. NXG Logic algorithms are comprised of a multitude of machine learning techniques, which can offer numerous perspectives on the results obtained. Computational intelligence and swarm intelligence algorithms are also employed to tackle a variety of numerical challenges.

Upgradability. NXG Logic offers several collections of toolkit availability, ranging from the Standard version, Professional version, to the Enterprise version.



Knowledge discovery and text mining. NXG Logic's technology empowers analysts to transform and transpose data, perform feature selection via cross-validation, dimension reduction with linear and non-linear manifold learning, cluster validity to determine the optimal number of clusters within a dataset, and assess association and generate force plots. NXG Logic's technology also includes text mining via stemming/stopping and N-gram analysis in order to cluster documents, sentences, abstracts, product reviews, and user comments. Extraction of concept clusters also includes graphic output to identify the key words (and documents) which are driving each concept.

Statistical analysis. Perform 2- and k-sample hypothesis testing, multiple linear regression, multivariate regression, polytomous (multi-class) logistic regression, Poisson regression, Cox proportional hazards regression, Kaplan-Meier analysis. Regression diagnostics are available for most of the regression models.

Component subtraction, decorrelation, denoising, and super-resolution root MUSIC. Perform component subtraction to decorrelate and denoise a dataset in order to reduce strong correlations and reduce uncertainties. Numerous time-consuming computational steps involving principal components analysis, multivariate linear regression, and fitting the Marčenko-Pastur limit distribution of eigenvalue density have been combined to automatically provide results. Covariance matrix filtering and super-resolution root MUSIC analysis are also available to reduce bias among data.

Dimension reduction and class discovery. NXG Logic has also developed numerous dimension reduction algorithms based on linear methods and non-linear manifold learning. Examples include correlation-based PCA (CPCA), kernel distance-based PCA (KDPCA), kernel Gaussian radial basis function PCA (KGPCA), kernel Tanimoto distance-based PCA (KTPCA), diffusion maps (DM), localized linear embeddings (LLE), Laplacian eigenmaps (LEM), and locally preserved projections (LPP). Identification of novel clusters in data can be determined by using NXG Logic's suite of class discovery tools, including crisp K-means cluster (CKM), fuzzy-K-means cluster (FKM), self-organizing maps (SOM), unsupervised neural gas (UNG), unsupervised artificial neural networks (UANN), Gaussian mixture models (GMM), unsupervised random forests (URF), Sammon mapping (Sammon), classic multidimensional scaling (CMDS), non-metric multidimensional scaling (NMMDS), and hierarchical cluster analysis (HCA).

Class prediction, cross validation, performance. NXG Logic's class prediction techniques include linear regression (LREG), decision tree classification (DTC), supervised random forests (SRF), K-nearest neighbor (KNN), naive Bayes classifier (NBC), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), Fisher discriminant analysis (FDA), learning vector quantization (LVQ), polytomous logistic regression (PLOG), gradient ascent support vector machines (SVMGA), least squares support vector machines (SVMLS), artificial neural networks (ANN), kernel regression (KREG), particle swarm optimization, supervised neural gas (SNG), and mixture of experts (MOE). The potential for selection bias can be minimized when performing class prediction by employing several types of cross-validation, such as bootstrap bias, k-fold, and leave one out (LOOCV). The performance of class prediction analysis can be evaluated for sensitivity/specificity, kappa vs. error (classifier diversity), receiver operator characteristic (ROC) curves, ROC area under the curve comparisons for all pairwise 2-class comparisons, as well as average area under the curve (AUC).

Simulation. NXG Logic technology also includes fitting probability distributions to data and simulating quantiles from normal, log-normal, chi-squared, Erlang, gamma, Student's t, F-ratio, Cauchy, Laplace, logistic, beta, betaPERT, Pareto, power, Rayleigh, triangle, stable distributions, and performing Monte Carlo uncertainty analysis based on correlated data. Monte Carlo cost analysis can also be performed using stored run parameters for distributions, correlations, etc.