

Data Mining and the Euredit Project

By Stephen Langdell, Numerical Algorithms Group

By definition, practitioners in multidisciplinary fields such as bioinformatics need to keep abreast of the latest research trends in several areas. However, the relevance of research work focused in areas outside their normal specialty may not be immediately obvious to all observers.

For instance, a multi-million dollar, 4-year research project on statistical methods for data mining, called Euredit and funded by the European Union, is creating interesting possibilities for many bioinformatics endeavors, including microarray analysis and forecasting trends.

Euredit's goal was to analyze census surveys. Due to human nature, these surveys usually contain missing or incorrect data. European governments funded research to determine and improve statistical techniques that can be used to clean this data by filling in gaps or highlighting errors. This research was carried out by government national statistics offices, universities, and research-based companies. The cleaned data produced by this research was then used to better estimate the need for services and programs in communities. The result of this intensive research study was a number of new algorithms relevant to disciplines where data mining is important, with bioinformatics perhaps one of the top beneficiaries.

For example, whether one is attempting to identify population subsets with certain characteristics as part of demographic analyses or looking for interesting groups of genes in microarray studies, the application of cluster analysis techniques is key. Cluster analysis functions in commercial packages require storage in computer memory of an n -by- n matrix of similarities (or differences) between " n " genes. This storage requirement limits the number of genes that can be studied at any one time.

Similar restrictions on dataset size were problematic for demographics researchers using cluster analysis techniques for European population analyses. Consequently, methods that obviated the need to store n -by- n similarity matrices were developed during the Euredit project, and hence provided the ability to study much larger datasets. These developments could equally be used in the bioinformatics field.

A similar Euredit development concerned logistic regression techniques. In bioinformatics logistic regression is used to classify data such as correctly assigning patients to risk levels or performing amino acid coding in DNA analyses. Here also, the typical size of datasets cannot readily be analyzed using traditional algorithms and hence can make for challenging research problems. To eliminate this problem, regression models with out-of-core optimization,

sometimes called data chunking, were developed during the Euredit project, dispensing with the need to store entire datasets in computer memory.

Another example of new data mining techniques developed during the Euredit project that can be applied in bioinformatics research are methods for identifying unusual cases hidden in data, known in data mining terms as outliers. After four years' study of various algorithms to handle outliers, the Euredit project developed methods to identify outliers in both categorical and continuous data. The algorithms were designed to handle very large data sets, and yield results with higher accuracy. These new data cleaning algorithms have yet to be applied extensively in bioinformatics research.

Until Euredit, decision trees -- methods used to discover diagnostic rules (i.e., rules in human-readable form) in data -- were very susceptible to outliers. To counter this, a regression tree was developed by Euredit that is robust with respect to outliers. It differs from all other regression trees by automatically weighting the data at nodes in a decision tree such that outlier effects are either removed or minimized.

At the conclusion of the Euredit project, the worldwide Numerical Algorithms Group (NAG) undertook to disseminate Euredit's findings. To that end NAG created the first commercially available data mining application toolkit that uses the new algorithms. These algorithms (along with many others) are provided as components that can be easily incorporated into user's existing applications.

To learn more about the functionality in NAG's new data mining components, please see <http://www.nag.com/numeric/DR/Drfunctionality.asp>