

BY RAY SOMORJAI, PhD

Twin Curses Plague Biomedical Data Analysis

Noninvasive experimental techniques, such as magnetic resonance (MR), infrared, Raman and fluorescence spectroscopy, and more recently, mass spectroscopy (proteomics) and microarrays (genomics) have helped us better understand, diagnose and treat disease. These methods create huge numbers of features, on the order of 1,000 to 10,000, resulting in Bellman's *curse of dimensionality*: too many features (i.e., dimensions). However, clinical reality frequently limits the number of available samples to the order of 10 to 100. This leads to the *curse of dataset sparsity*: too few samples. Thus, on the one hand, we have a wealth of information available for data analysis; on the other hand, statistically meaningful analysis is hampered by sample scarcity.

Robust, reliable data classification (e.g., distinguishing between diseased and healthy conditions) requires a sample-to-feature ratio on the order of 5 to 10, instead of the initial 1/10 to 1/1000. What can be done?

To lift the curse of dimensionality and reduce the number of features to a manageable size, we use feature extraction/selection (FES). FES reduces dimensionality by identifying and eliminating redundant or irrelevant information. For microarray data this is accomplished by first identifying groups of correlated genes and

defining group averages as new features. For spectra, neighboring features are strongly correlated, and therefore the majority of features are redundant. In addition, many features are "noise," or are irrelevant for the desired classification. Eliminating these yields a much lower-dimensional feature space that suffices for accurate spectral characterization. To identify the spectral fea-



Dataset sparsity has more subtle consequences, and lifting this curse is more problematic. The ideal solution—acquiring more samples—is frequently too expensive or even impracticable.

tures to be eliminated, we have developed an algorithm that helps select optimal sub-regions that are most relevant for an accurate classification. Averaging adjacent spectral intensities leads to further reduction, while retaining spectral identity, which is important for interpretability of the resulting features (e.g., MR peaks, essentially averages of adjacent spectral intensities, are manifestations of the presence of specific chemical compounds).

Dataset sparsity has more subtle consequences, and lifting this curse is more problematic. The ideal solution—acquiring more samples—is frequently too expensive or even impracticable. Yet, limited sample size may create classifiers that give overoptimistic accuracies, even after feature space reduction. Robust classifier creation requires enough samples to meaningfully partition the data into training, validation and independent test sets. The training set is used for both FES and optimal classifier development. The validation set helps prevent the classifier from adapting to the peculiarities of a finite train-

DETAILS

Ray Somorjai is Head of Biomedical Informatics at the Institute for Biodiagnostics, National Research Council Canada. The three major thrusts in his Group are supervised classification, with special emphasis on handling the peculiarities of biomedical data, unsupervised classification (e.g., EvIdent, a powerful fuzzy clustering software) and the mathematical modeling of the spread of infectious diseases (AIDS, SARS, etc.).

continued on page 25

continued from page 24

ing set (overfitting) by monitoring the progress of the FES/classifier. The independent test set is used for external cross-validation, but only after completion of the FES and identification of the final classifier. With small datasets, even partitioning into training and test sets is statistically suspect, and k -fold cross-validation is used: the dataset is split into k equal parts (approximately 5 to 10), trained on $k-1$ parts and tested on the remaining portion. One then cycles through k times and averages the test results. For small sample sizes, the variance of the averaged test accuracies tends to be unacceptably large, while over-training is still a threat.

For highly imbalanced classes (e.g., rare disease vs. healthy), overall classification accuracy can be misleading. For example, consider 90

samples in the healthy class, but only 10 in the disease class. Misclassifying all 10 still gives 90 percent overall accuracy. Hence, balanced sensitivity and specificity (i.e., comparable accuracies for both classes) is more appropriate, and can be achieved by undersampling, oversampling or by penalizing misclassifications differently for different classes. (Differing misclassification costs for the classes is an example.)

For each sample, we compute class probabilities. This is relevant clinically (e.g., additional tests would be suggested if a classifier assigned a patient to the disease class with 55 percent probability, immediate treatment would commence if this probability were 90 percent.)

In the biomedical field, the twin curses are generally active. They both must be dealt with in concert, otherwise overly optimistic and frequently wrong conclusions will result. □

SeeingScience continued

continued from page 26

With a name inspired by Friedrich Nietzsche's *Ecce Homo*, a meditation on how one becomes what one is, the project explores human evolution by examining similarities between genes from human beings and a target organism, in this case the rice plant. *Ecce*

Ecce Homology is a physically interactive new-media work that visualizes genetic data as calligraphic forms.

Homology is a physically interactive new-media work that visualizes genetic data as calligraphic forms. A novel computer-vision based interface allows multiple participants, through their movement in the installation space, to select genes from the human genome for visualization using the Basic Local Alignment Search Tool (BLAST). Five projectors present these changes in *Ecce Homology's* calligraphic forms across a 40-foot wide wall.

"If we worked on the genomic cal-

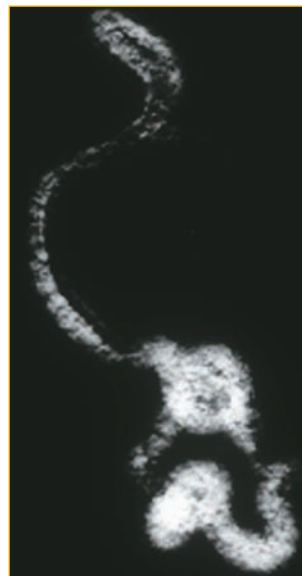
ligraphy visualization further, it could be useful to scientists," she says, "but the installation is not a tool; it's art. And it's specifically ambiguous and a bit mysterious—by intention."

Ecce Homology, which was first displayed two years ago at the Fowler Museum in Los Angeles, works on many levels both scientifically and artistically. "People assume that there's value in the vast amounts of genomic data we are generating," says West, "but data is not knowledge, and in order for us to derive knowledge from it, we need to interpret it. The more complex it is, the harder it is for human beings to do that and, consequently, the greater our need to find new approaches." So, says West, "we've produced an artwork that both speaks to this need and lets viewers interact fluid-

ly with the data in a visceral way."

Ultimately, West says, the exhibit poses the question, "If you were to do work that's truly hybrid art/science, what would that process be like? And would there be any outcome that would point to how art might nurture scientific discovery?"

For more information about *Ecce Homology*, visit www.insilicov1.org. □



***Ecce Homology's* custom software transforms strings of genetic code into luminous, scientifically accurate visualizations that incorporate multiple biological features. For protein sequences, the stroke placement, shape and brush quality are determined by physical and chemical properties, such as the proportion of mass to volume, hydrophobicity, or ionization of the amino acids. The visualization is created from amino-acid sequence chunks that are segmented by a "turn prediction" algorithm. Each segment's corresponding calligraphic stroke is connected to its neighbor by a connection whose shape is based on a secondary structure property of the segment. The result resembles calligraphy. Courtesy: Ruth West**