Brown mathematicians prove new way to build a better estimate

February 29 2008

How do you sift through hundreds of billions of bits of information and make accurate inferences from such gargantuan sets of data? Brown University mathematician Charles “Chip” Lawrence and graduate student Luis Carvalho have arrived at a fresh answer with broad applications in science, technology and business.

In new work published in the *Proceedings of the National Academy of Sciences*, Lawrence and Carvalho describe a new class of statistical estimators and prove four theorems concerning their properties. Their work shows that these “centroid” estimators allow for better statistical predictions – and, as a result, better ways to extract information from the immense data sets used in computational biology, information technology, banking and finance, medicine and engineering.

“What’s exciting about this work – what makes it every scientist’s dream – is that it’s so fundamental,” Lawrence said. “These new estimators have applications in biology and beyond and they advance a statistical method that’s been around for decades.”

For more than 80 years, one of the most common methods of statistical prediction has been maximum likelihood estimation (MLE). This method is used to find the single most probable solution, or estimate, from a set of data.

But new technologies that capture enormous amounts of data – human genome sequencing, Internet transaction tracking, instruments that beam high-resolution images from outer space – have opened opportunities to predict discrete “high dimensional” or “high-D” unknowns. The huge number of combinations of these “high-D” unknowns produces enormous statistical uncertainty. Data has outgrown data analysis.

This discrepancy creates a paradox. Instead of producing more precise predictions about gene activity, shopping habits or the presence of faraway stars, these large data sets are producing more unreliable predictions, given current procedures. That’s because maximum likelihood estimators use data to identify the single most probable solution. But because any one data point swims in an increasingly immense sea, it’s not likely to be representative.

Lawrence, a professor of applied mathematics and a faculty member in the Center for Computational Molecular Biology at Brown, first came upon this paradox and a potential way around it while working on predicting the structure of RNA molecules. If you want to predict the structure of these molecules – how the molecule will look when it folds onto itself – you’d have billions and billions of possible shapes to choose from.

“Using maximum likelihood estimation, the most likely outcome would be very, very, very unlikely,” Lawrence said, “so we knew we needed a better estimation method.”

Lawrence and Carvahlo used statistical decision theory to understand the limitations of the old procedure when faced with new “high-D” problems. They also used statistical decision-making theory to find an estimation procedure that applies to a broad range of statistical problems. These “centroid” estimators identify not the single most probable solution, but the solution that is most representative of all the data in a set.

Lawrence and Carvahlo went on to prove four theorems that illustrate the favorable properties of these estimators and show that they can be easily computed in many important applications.

“This new procedure should benefit any field that needs to reliably make predictions of large-scale, high-D unknowns,” Lawrence said.

Source: Brown University

Citation: Brown mathematicians prove new way to build a better estimate (2008, February 29) retrieved 30 April 2024 from https://phys.org/news/2008-02-brown-mathematicians.html