25 years of conventional evaluation of data analysis proves worthless in practice

September 3 2008

So-called 'intelligent' computer-based methods for classifying patient samples, for example, have been evaluated with the help of two methods that have completely dominated research for 25 years. Now Swedish researchers at Uppsala University are revealing that this methodology is worthless when it comes to practical problems. The article is published in the journal *Pattern Recognition Letters*.

Today there is rapidly growing interest in 'intelligent' computer-based methods that use various classes of measurement signals, from different patient samples, for instance, to create a model for classifying new observations. This type of method is the basis for many technical applications, such as recognition of human speech, images, and fingerprints, and is now also beginning to attract new fields such as health care.

"Especially in applications in which faulty classification decisions can lead to catastrophic consequences, such as choosing the wrong form of therapy for treating cancer, it is extremely important to be able to make a reliable estimate of the performance of the classification model," explains Mats Gustafsson, Professor of signal processing and medical bioinformatics at Uppsala University, who co-directed the new study together with Associate Professor Anders Isaksson.

To evaluate the performance of a classification model, one normally tests it on a number of trial examples that have never been involved in the design of the model. Unfortunately there are seldom tens of

thousands of test examples available for this type of evaluation. In biomedicine, for instance, it is often expensive and difficult to collect the patient samples needed, especially if one wishes to analyze a rare disease. To solve this problem, many different methods have been proposed. Since the 1980s two methods have completely dominated research, namely, cross validation and resampling/bootstrapping.

"This has entailed that the performance assessment of virtually all new methods and applications reported in the scientific literature in the last 25 years has been carried out using one of these two methods," says Mats Gustafsson.

In the new study, the Uppsala researchers use both theory and convincing computer simulations to show that this methodology is worthless in practice when the total number of examples is small in relation to the natural variation that exists among different observations. What is considered a small number depends in turn on what problem is being studied-in other words, it is impossible to determine whether the number of examples is sufficient.

"Our main conclusion is that this methodology cannot be depended on at all, and that it therefore needs to be immediately replaces by Bayesian methods, for example, which can deliver reliable measures of the uncertainty that exists. Only then will multivariate analyses be in any position to be adopted in such critical applications as health care," says Mats Gustafsson.

Source: Uppsala University