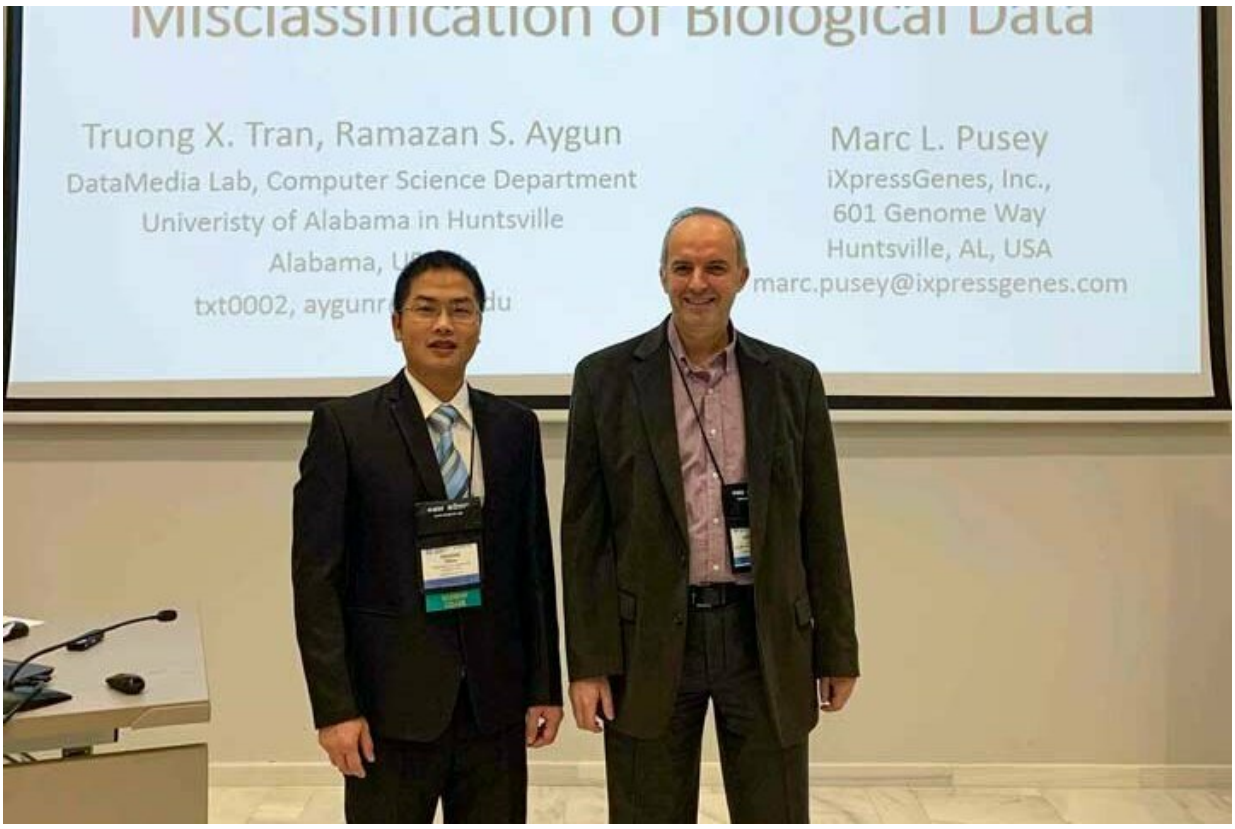
Novel else-tree classifier seeks to minimize misclassification in biological research studies

January 29 2019, by Diana Lachance



The research conducted by Tran and his advisor, Dr. Aygun, on the else-tree classifier is part of a larger project entitled “Macromolecule Crystallization Screening Results Analysis,” which is being funded by a grant from the National Institutes of Health and sponsored by iXpressGenes. Credit: Truong Xuan Tran

Truong Xuan Tran may have started his academic career as an electronic and telecommunication engineering major at the Hanoi University of Science and Technology, but since coming to the U.S., he's decided to follow his passion for computer science instead. In 2012, the Vietnam native earned his master's degree in the field from Arkansas State University, and in 2016, he joined the Department of Computer Science at The University of Alabama in Huntsville (UAH) in pursuit of his Ph.D. "My research interests are machine learning and data mining," says Tran.

For the last two years, he has been working closely with Dr. Ramazan Aygun, an associate professor in UAH's Department of Computer Science and principal investigator on a project entitled "Macromolecule Crystallization Screening Results Analysis." The project is funded by a grant from the National Institutes of Health and sponsored by iXpressGenes, a biotechnology company specializing in protein services and instrumentation that was founded by UAH biology professor Dr. Joseph Ng.

"Protein crystallization is a difficult process where thousands of trials may need to be set up for a successful crystalline outcome," says Dr. Aygun, who co-authored the 2017 book "Data Analytics for Protein Crystallization" with iXpressGenes' senior scientist Mark Pusey. "Data analytics include methods for providing conditions to be set up, automated scoring of images, protein growth analysis, focal stacking for trial images, and visualization of plates."

In December, Tran and Dr. Aygun traveled to Madrid, Spain, for the IEEE International Conference on Bioinformatics and Biomedicine, where they presented their most recent research results at the workshop on Machine Learning and Artificial Intelligence in Bioinformatics and Medical Informatics. Their paper, co-written with Dr. Pusey and entitled "Else-Tree Classifier for Minimizing Misclassification of Biological

Data," introduced a novel decision-tree classifier, dubbed else-tree, that reduces the misclassification of data samples by labeling them as undecided rather than assigning them an incorrect class. "The main feature of the else-tree is its potential to generate zero percent error without overfitting by separating hard-to-classify data as undecided," says Tran.

As he explains, this stands in contrast to traditional decision tree classifiers, which are based on an impurity measure that identifies the most informative attribute to be selected at the early levels of a decision tree. In many cases, the classification model yielding highest evaluation value is then selected to become the predictor for future data; however, a significantly high value of an evaluation measure is an indication of overfitting, and so the classifiers most likely to be chosen are also likely to have false classifications for new unseen data.

"In biological research studies such as protein crystallization, inaccuracy or misclassification of machine learning algorithms can yield fatal results," says Tran. For example, if the 3-D structure of a protein is initially obtained by crystallizing the protein in drug development, missing a crystalline condition may hinder its development. "So it is important to develop classification algorithms that would avoid or minimize misclassifications."

With the else-tree, data that would otherwise be misclassified is instead labeled undecided and an additional virtual class is generated, which both avoids critical mistakes and increases the trust of the user of the classifier. "The key point of the else-tree is that it postpones difficult data to classify by sending them to its else branch until a good attribute can classify those samples," says Tran. "We have also introduced a pruning method based on the number of elements in a region, wherein the data are considered for classification and labeled if they have some minimum number of elements."

Experimental results on three public data sets and the team's protein crystallization data have already shown that the else-tree can outperform other methods when the training data is a representative of the complete set. And they are hoping to improve on these results going forward, with an eye toward minimizing or avoiding errors that could happen at early branches of the tree. "We plan to investigate the proposed algorithm on other types of problems such as multi-class classification," says Tran. "We will examine the effectiveness of our method on other datasets. Especially, we will work on improving the else-tree to reduce the percentage of undecided samples while keeping the error minimum by working on other datasets."

Provided by University of Alabama in Huntsville

Citation: Novel else-tree classifier seeks to minimize misclassification in biological research studies (2019, January 29) retrieved 19 April 2024 from https://phys.org/news/2019-01-else-tree-minimize-misclassification-biological.html