University of Cyprus
Department of Computer Science

You are cordially invited to the PhD Defense on:

Machine Learning for Privacy Preserving Data Publishing and the Analysis of Categorical Data in the Medical Domain

Mr. Aristos Aristodimou
University of Cyprus, Cyprus

Thursday, November 28, 2019
10:00-11:00 EET

Room 148, Building 12
Faculty of Pure and Applied Sciences, New Campus

Abstract

In recent years, with the infiltration of information technology in healthcare, many healthcare related entities, have vast amounts of patients' data. Although sharing such data can increase the likelihood of identifying novel findings or even replicating existing research results, this is not happening due to legal and ethical issues. Machine learning can be used on such datasets to identify risk factors that can be used to improve our lives, but any algorithms that will analyze such data should take into consideration that most common diseases are influenced by multiple gene interactions and interactions with the environment. Hence they should use models that allow the finding of such multivariate associations. This work initially presents a novel algorithm for anonymizing categorical data with k- anonymity and performing feature selection for classification tasks. The algorithm was evaluated on various medical datasets and in the majority of the evaluated test cases the produced anonymized data had similar or better accuracies than using the full datasets. Additionally, a novel density based discretization algorithm is presented that has similar performance with state of the art algorithms while being computationally efficient and suitable for big data. For pattern recognition of n-SNP associations in case/control data, a Self Organizing Map for nominal categorical data is presented, which was able to produce statistically significant clustering revealing some interesting patterns between the clusters of cases and controls. Finally, a framework for efficient n-Way interaction testing is presented that uses machine learning to reduce the dimensionality of the data and to produce a targeted binary encoding of the genotypes. This enables the reduction of the multiple testing problem and the degrees of freedom of the statistical tests applied for interaction testing, and hence increases the statistical power of the performed analysis. Results indicate that with the new encoding, the proposed framework was able to identify more statistically significant interactions compared to using the initial encoding of the genotypes.

Short Bio:

Aristos Aristodimou is a Ph.D. candidate at the Department of Computer Science under the supervision of Professor Constantinos S. Pattichis. His research interests lie in the area of machine learning and its application in the medical domain. He received his B.Sc. in Computer Science from the University of Cyprus and his M.Sc. in Informatics – Learning from Data from the University of Edinburgh.

Host: Prof. Chris Christodoulou (cchrist-AT-cs.ucy.ac.cy)