ST. FRANCIS XAVIER UNIVERSITY

Department of Mathematics, Statistics and Computer Science

St. Francis Xavier University

Presents

Classifying and Clustering Drug Discovery Data

by

Fateha Khanam Bappee

St. Francis Xavier University

M.Sc. Thesis Proposal Presentation

Thursday, July 18 @ 10:00am in Annex 23A

A significant development of computer based technology and bioinformatics has motivated the radical change in drug discovery. The goal of drug discovery is to identify active compounds against biological target. This thesis uses AIDS antiviral screening data set from the National Cancer Institute (NCI) Developmental Therapeutic Program. The two main parts of the thesis are: classifying and clustering drug discovery data. We propose an extension of the regular K-Nearest Neighbour (KNN) method named Adaptively Chosen *k* in Weighted KNN (Ad-WKNN). The main features of Ad-WKNN are choosing adaptive *k* and employing a weighting scheme for the nearest neighbours according to their similarity to a test point; as well as handling a mixed type of descriptors simultaneously. Our preliminary results show that the proposed Ad-WKNN method performs better than the regular KNN method in predicting more active compounds. In the second part of the thesis, a popular clustering method, K-means is used to group drug discovery data. We propose an evaluation measure to determine suitable values of K. Instead of using random initial centroids, the initial cluster centers are calculated based on the size of clusters. Furthermore, the method takes into consideration of the mixed types of variables.

Refreshments will be served before the talk in AX24A