**Title - Exploring automated Feature Selection for Model-based and Density-based clustering with application to NCI 60 data**

**Abstract**

A major objective of clustering is to identify groups in the data that maximizes the similarity between objects within the same cluster and minimizes the similarity between different clusters. A challenge for data clustering, and unsupervised learning in general, is that there is often no mechanism for feature selection, unlike the supervised learning problems which can be solved in connection with the subset selection or LASSO like penalties. In contrast, variable selection in unsupervised learning problems is not well defined as there is no response variable and the subset selection is far more challenging. Consequently, there has been comparatively few methods that automate feature selection for clustering. Typically, when faced with high dimensionality, or the possibility of irrelevant features, an investigator will employ dimension reduction techniques with standard clustering algorithms. In this work, we explored two principal methods that encode feature selection into the clustering process, Model based clustering and Density based clustering using the Clustvarsel and DBSCAN packages were used to show the working of these clustering techniques. These methods were applied to the NCI-60 data, and compared to the principal component based k-means over different parameter settings. In conclusion, the performance of k-means algorithm is better than Density Based clustering algorithm and the Model Based clustering algorithm. The performance of DBSCAN is more efficient and accurate, and is less time taking when compared to the working of the Clustvarsel algorithm.

Suruchi Jai Kumar Ahuja

MS – Bioinformatics and Biostatistics.