**Activities and Findings Related to NSF grant**

Small:Simultaneous Decomposition and Predictive Modeling on Large Multi-Modal Data

**Research and Education Activities:**

1. Espoused the philosophy of Simultaneous Decomposition and Prediction (SDaP) to a broad audience (keynote, book chapter, papers). Also gave invited keynote talks abroad (Spain, Brazil, Mexico), in addition to several domestic forums.

2. Formulated a hard decomposition version of SDaP that includes model selection (number of decomposed pieces) and devised new, effective active learning principles for a hard version of SDaP

3. Reviewed the literature on combining multiple clusterings to provide a conceptual hierarchy on the diverse ideas (review paper published, two book chapters in the works), and then formulated a novel framework called C3E that combines classifier ensembles and cluster ensembles to deal with using both labelled and (subsequent) unlabelled data, even when the underlying models change over time. Showed its power via extensive experiments. Journal version of C3E (to appear in IEEE Trans. TKDE). Also devised an ensemble based approach to the imbalanced class problem (when one class is very rare) using alpha divergence. Journal paper appeared in IEEE Trans. KDE.

4. Bayesian formulation of soft SDaP using generative models completed with full variational inference. Subsequently we generalized this approach to one of Constrained Relative Entropy Minimization with Applications to Multitask Learning. This has quite well developed, and the core of Koyejo's PhD thesis, which he successfully defended in May 13. Several publications have already resulted (including one that received Amazon's Best Student Paper at UAI'13), and a journal paper accepted in Machine Learning Jl.

5. The problem of ranking on networks is closely related to this project. We have developed a set of tools using ideas of monotonicity and covexity, with results that are beating other ranking methods such as CofiRank. This work has led to Acharyya's PhD thesis (graduated Aug 13), and papers in UAI12, Recsys13 and UAI14.

6. Fruitful collaboration with Yahoo, with visits to both Mountainview and Barcelona, has resulted. They have provided two very large datasets, but with restrictions on publication, specially one that deals with actual customer data from their Taiwanese properties. However these datasets give us an understanding of real life scale and data quality issues.

7. I offered a 1 month course on advanced data mining at UNICAMP, Brazil (ranked #2 in engineering/sciences in Latin America), with over 6 hours of lectures devoted to ideas and results intimately related to this project.

8. We have recently applied SDAP concepts to high throughput phenotype

extraction from large scale EHR data, and our first work will appear in Jl. Biological

Informatics.