香港浸會大學理學院
HKBU Faculty of Science

for favour of posting

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Mr Hong ZENG
Supervisor:	Dr Yiu Ming CHEUNG
External Examiner:	Prof Ping GUO Dr Hon Wah TAM
Time:	18 September 2009 (Friday) 10:30 am – 12:30 pm (35 mins presentation and 15 mins Q & A)
Venue:	T909, Cha Chi Ming Science Tower, HSH Campus

"On Feature Selection, Kernel Learning and Pairwise Constraints for Clustering Analysis"

Abstract

Given an objective function and a metric measure, the performance of most clustering algorithms strongly depend on the data representation. The first objective of this thesis is then to study how to learn a concise and informative representation for clustering analysis through feature selection and kernel learning. Besides, since the many clustering algorithms may not be able to produce user-desired groups, the second objective of this thesis is to investigate how to effectively incorporate user-provided pairwise constraints, which specify whether the pair of point should be in the same cluster or not, to assist the clustering.

Regarding to the first objective, we propose two feature selection methods as well as a kernel learning method. The first feature selection algorithm is proposed to reduce the dimension of data before performing clustering analysis. The importance of each feature is evaluated by its locality preserving capability, only those with higher locality preserving power are selected. To better characterize the local geometric structure, we formulate the locality preserving criterion based on the local kernel ridge regression method which is effective in modeling the local variation of target values. Experimental results on several benchmark datasets demonstrate that this method can maintain a low-dimensional representation for the data effectively. The second feature selection method is proposed to improve the performance of a specific clustering algorithm. This clustering algorithm employs the local learning idea to derive a clustering objective function which enforces the local smoothness on cluster labels. In the local learning, a local model is trained in the

neighborhood of each sample, exclusively of samples out of the neighborhood. Therefore, it may suffer from the insufficient number of samples in the neighborhood and the presence of many irrelevant features. We formulate a training scheme under the multi-task learning framework, in which local models are coupled by sharing a nonnegative feature weight vector among them. Such a paradigm is theoretically guaranteed to produce a sparse weighting where the weight for irrelevant feature is driven towards zero (which corresponds to performing feature selection), as well as sparse local models in which only a small number of informative features are involved. Thereby, the local smoothness criterion can be refined with such sparse weighting, leading to better partitioning. Furthermore, a kernel learning algorithm is derived from the second feature selection method using the kernel trick, where a convex combination of kernels is learned for the clustering. The effectiveness of the proposed approaches is validated on extensive real-world datasets.

For the second objective of this thesis, we propose to incorporate pairwise constraints into a clustering algorithm which finds maximum margin hyperplanes to separate the data. We further introduce a set of loss functions for effectively penalizing the violation of the given pairwise constraints. Unlike previous pairwise constrained clustering approaches which attempt to learn better distance metrics or improve the estimation for the underlying data distribution, the proposed approach aims at finding better clustering boundaries, thus requiring fewer model assumptions. The resulting problem is solved by the Constrained Concave-Convex Procedure (CCCP), with an efficient subgradient projection routine embedded. Comparison of the proposed method with several competitive approaches over a number of benchmark real-world datasets, shows that the proposed method is more computational efficient and able to propagate the pairwise constraints more effectively. Last but not the least, it also has good generalization performance on out-of-sample points.

***** ALL INTERESTED ARE WELCOME *****