Large Scale Metric Learning for Matching of Heterogeneous Multimedia Data

(PI: Prof. Zhang Dapeng David; 2014/15)

Heterogeneous multimedia data are widely encountered in many applications, such as photo-sketch face recognition, still image to video face recognition, cross-modality image synthesis, cross media retrieval, etc. With the ubiquitous use of digital imaging devices, mobile terminals and social networks, there are lots of heterogeneous and homogeneous data from multiple sources, e.g., news media websites, microblog, mobile phone, social networking, etc. Matching of heterogeneous multimedia data becomes increasingly important to achieve cross modal and cross media information retrieval. One popular approach to the matching of heterogeneous data is metric learning, which learns a positive semi-definite matrix to measure the similarity of heterogeneous data. However, there are several challenging issues that the current metric learning methods cannot adequately address. First, the current metric learning methods are limited in dealing with the highly diverse and complex data types in real-world, including text, graph, image, audio, video, 2D and 3D data, etc. Second, the current methods have poor scalability, which is a critical issue in handling the tremendous amount of multimedia data. Third, the labels of most data are unavailable, making them difficult to be used by current metric learning methods. Fourth, the data in the same modality may have different representations, and thus a multiple feature

similarity metric should be learned for cross modal data matching. Hence, it is highly desirable to develop a new metric learning model for cross modal and cross media data matching, which can have good scalability, handle effectively unlabeled data, and measure multiple feature similarity across modals.

In this project, we propose a large-scale pairwise kernel classification model to learn cross modal distance metrics for matching of heterogeneous multimedia data. We formulate the cross modal metric learning problem as a sample pair classification problem with pairwise constraints from training samples, and then develop support vector machine (SVM) based iterated training algorithms to efficiently solve the model. The proposed metric learning model has the following advantages. First, with the defined cross modal pairwise kernel, data-dependent kernels, e.g., Gaussian kernel and histogram kernel, can be introduced to deal with the complex types of data. Second, cache techniques will be introduced, which makes it feasible to process millions of pairwise constraints. By the proposed iterated SVM training algorithms, large-scale metric learning can be conducted to learn desired cross-modal distance metrics from huge data. Third, graph Laplacian regularization can be introduced to our metric learning model to utilize the abundant unlabeled or linked data, and then the semi-supervised model can be solved by Laplacian-SVM. Fourth, considering that in each modal the data may have different representations, we propose a multiple kernel metric

learning model to effectively combine the different representations.

The proposed metric learning model can boost significantly the matching performance of heterogeneous multimedia data, and has a wide range of applications in large scale cross-media retrieval. Our preliminary results are very encouraging, validating the feasibility of this project and its great potential. The research outputs of this project will initiate new directions of cross-modal data matching, and will have high impact in the fields of multimedia content analysis and metric learning.