Abstract

The rapid development of information and computer technology (ICT) in the last two decades has fundamentally changed almost every discipline in science and engineering, transforming many fields from data-poor to increasingly datarich, and calling for innovative data mining methods to conduct the related research. Meanwhile, as data collection sources and channels continuously evolve, data can be extracted from multiple information sources and observed by various models. Therefore, learning from multi-view data has become a crucial step in machine intelligence and knowledge discovery.

For the purpose of integrating and leveraging the mass amount of multi-view data to obtain significant and complementary high-level knowledge, this dissertation investigates learning from multi-view data from two sides: clustering algorithm and text mining application.

The dissertation is organized into three parts. In the first part, we analyze multi-view clustering from a multilinear perspective and create several novel multi-view clustering algorithms. At first, modeling multi-view data as a tensor, we present a novel tensor based multi-view partitioning framework for integrating multi-view data in the context of spectral clustering. Within this framework, a joint optimal subspace shared by multi-view data as well as the multilinear relationships among multi-view data are revealed by the relevant tensor methods.

Second, taking multi-view data as multiple graphs, we put forward a multi-view clustering strategy based on simultaneous trace maximization (STM), which analyzes multi-view data through a multilinear perspective as well. Third, a joint dimension reduction scheme based on tensor decomposition is presented, particularly for multi-view data. The dimension reduction scheme is embedded into the STM based multi-view clustering strategy, which enables us to handle large-scale multi-view data. In the second part, we investigate text mining to extract multi-view heterogeneous data from a large-scale publication database of Web of Science (WoS).
In order to facilitate the scientific mapping that is useful for monitoring and detecting new trends in different scientific fields, hybrid clustering, either in vector spaces or in graph spaces, is carried out to integrate these multi-view data. Regarding hybrid clustering in vector spaces, various methodologies are included in a unified framework, which consists of two general approaches: clustering ensemble and kernel fusion. A mutual information based weighting scheme is proposed to leverage the effect of multiple data sources in hybrid clustering. Concerning hybrid clustering in graph spaces, various graphs are generated from multi-view data. Utilizing the complementary properties of both text graph and citation graph, we present a hybrid strategy named graph coupling. Meanwhile, based on the modularity optimization, our graph coupling strategy detects the number of clusters automatically and provides a top-down hierarchical analysis, which fits in with the practical applications. In addition, the computation of this modularity based hybrid clustering method is so efficient that it does well in partitioning large-scale data.

In the third part, we propose a novel strategy to derive knowledge from textual information from a multi-view perspective. The multiple views can be different controlled vocabularies, term weighting schemes, publishing time periods and biomedical subjects. Our strategy has been applied to the MEDLINE corpus and analyzed using a disease based data set. In particular, we investigate the effect of combining multiple views for clustering and assessed whether vertical searches can be more accurate for specific biological questions. Moreover, a Web application of our multi-view text mining strategy is developed for gene retrieval.

To conclude, the theory, algorithm, applications and software presented in this dissertation provide an interesting perspective for clustering algorithms and text mining applications. In addition, the obtained results are promising to be applied and extended to many other relevant fields besides scientific mapping and bioinformatics.