Abstract

The emerging problem of data fusion offers plenty of opportunities, also raises lots of interdisciplinary challenges in computational biology. Currently, developments in high-throughput technologies generate Terabytes of genomic data at awesome rate. How to combine and leverage the mass amount of data sources to obtain significant and complementary high-level knowledge is a state-of-art interest in statistics, machine learning and bioinformatics communities.

Supervised and unsupervised learning are fundamental topics in statistics and machine learning. To incorporate various learning methods with multiple data sources is a rather recent topic. In the first part of the thesis, we theoretically investigate a set of learning algorithms in statistics and machine learning. We find that many of these algorithms can be formulated as a unified mathematical model as the Rayleigh quotient and can be extended as dual representations on the basis of Kernel methods. Using the dual representations, the task of learning with multiple data sources is related to the kernel based data fusion, which has been actively studied in the recent five years.

In the second part of the thesis, we create several novel algorithms for supervised learning and unsupervised learning. We center our discussion on the feasibility and the efficiency of multi-source learning on large scale heterogeneous data sources. These new algorithms are encouraging to solve a wide range of emerging problems in bioinformatics and text mining.

In the third part of the thesis, we substantiate the values of the proposed algorithms in several real bioinformatics and journal scientometrics applications. These applications are algorithmically categorized as ranking problem and clustering problem. In ranking, we develop a multi-view text mining methodology to combine different text mining models for disease relevant gene prioritization. Moreover, we solidify our data sources and algorithms in a gene prioritization software, which is characterized as a novel kernel-based approach to combine text mining data with heterogeneous genomic data sources using phylogenetic evidences across multiple species. In clustering, we combine multiple text mining models and multiple genomic data sources to identify the disease relevant partitions of

genes. We also apply our methods in scientometric field to reveal the topic patterns of scientific publications. Using text mining technique, we create multiple lexical models for more than 8000 journals retrieved from Web of Science database. We also construct multiple interaction graphs by investigating the citations among these journals. These two types of information (lexical /citation) are combined together to automatically construct the structural clustering of journals. According to a systematic benchmark study, in both ranking and clustering problems, the machine learning performance is significantly improved by the thorough combination of heterogeneous data sources and data representations.

The theory, algorithms, applications and software presented in the thesis provide an interesting perspective for kernel-based data fusion in bioinformatics. Moreover, the obtained results are promising to be applied and extended to many other relevant fields besides bioinformatics and text mining.