DOCTORAL THESIS

Regularized models and algorithms for machine learning

Shen, Chenyang

Date of Award:
2015

ABSTRACT

Multi-lable learning (ML), multi-instance multi-label learning (MIML), large network learning and random under-sampling system are four active research topics in machine learning which have been studied intensively recently. So far, there are still a lot of open problems to be figured out in these topics which attract worldwide attention of researchers. This thesis mainly focuses on several novel methods designed for these research tasks respectively.

Then main difference between ML learning and traditional classification task is that in ML learning, one object can be characterized by several different labels (or classes). One important observation is that the labels received by similar objects in ML data are usually highly correlated with each other. In order to exploring this correlation of labels between objects which might be a key issue in ML learning, we consider to require the resulting label indicator to be low rank. In the proposed model, nuclear norm which is a famous convex relaxation of intractable matrix rank is introduced to label indicator in order to exploiting the underlying correlation in label domain. Motivated by the idea of spectral clustering, we also incorporate information from feature domain by constructing a graph among objects based on their features. Then with partial label information available, we integrate them together into a convex low rank based model designed for ML learning. The proposed model can be solved efficiently by using alternating direction method of multiplier (ADMM). We test the performance on several benchmark ML data sets and make comparisons with the state-of-art algorithms. The classification results demonstrate the efficiency and effectiveness of the proposed low rank based methods.

One step further, we consider MIML learning problem which is usually more complicated than ML learning: besides the possibility of having multiple labels, each object can be described by multiple instances simultaneously which may significantly increase the size of data. To handle the MIML learning problem we first propose and develop a novel sparsity-based MIML learning algorithm. Our idea is to formulate and construct a transductive objective function for label indicator to be learned by using the method of random walk with restart that exploits the relationships among instances and labels of objects, and computes the affinities among the objects. Then sparsity can be introduced in the labels indicator of the objective function such that relevant and irrelevant objects with respect to a given class can be distinguished. The resulting sparsity-based MIML model can be given as a constrained convex optimization problem, and it can be solved very efficiently by using the augmented Lagrangian method (ALM). Experimental results on benchmark data have shown that the proposed sparse-MIML algorithm is computationally efficient, and effective in label prediction for MIML data. We demonstrate that the performance of the proposed method is better than the other testing MIML learning algorithms.

Moreover, one big concern of an MIML learning algorithm is computational efficiency, especially when figuring out classification problem for large data sets. Most of the existing methods for solving MIML problems in literature may take a long computational time and have a huge storage cost for large MIML data sets. In this thesis, our main aim is to propose and develop an efficient Markov Chain based learning algorithm for MIML problems. Our idea is to perform labels classification among objects

and features identification iteratively through two Markov chains constructed by using objects and features respectively. The classification of objects can be obtained by using labels propagation via training data in the iterative method. Because it is not necessary to compute and store a huge affinity matrix among objects/instances, both the storage and computational time can be reduced significantly. For instance, when we handle MIML image data set of 10000 objects and 250000 instances, the proposed algorithm takes about 71 seconds. Also experimental results on some benchmark data sets are reported to illustrate the effectiveness of the proposed method in one-error, ranking loss, coverage and average precision, and show that it is competitive with the other methods.

In addition, we consider the module identification from large biological networks. Nowadays, the interactions among different genes, proteins and other small molecules are becoming more and more significant and have been studied intensively. One general way that helps people understand these interactions is to analyze networks constructed from genes/proteins. In particular, module structure as a common property of most biological networks has drawn much attention of researchers from different fields. However, biological networks might be corrupted by noise in the data which often lead to the miss-identification of module structure. Besides, some edges in network might be removed (or some nodes might be miss-connected) when improper parameters are selected which may also affect the module identified significantly. In conclusion, the module identification results are sensitive to noise as well as parameter selection of network. In this thesis, we consider employing multiple networks for consistent module detection in order to reduce the effect of noise and parameter settings. Instead of studying different networks separately, our idea is to combine multiple networks together by building them into tensor structure data. Then given any node as prior label information, tensor-based Markov chains are constructed iteratively for identification of the modules shared by the multiple networks. In addition, the proposed tensor-based Markov chain algorithm is capable of simultaneously evaluating the contribution from each network. It would be useful to measure the consistency of modules in the multiple networks. In the experiments, we test our method on two groups of gene co-expression networks from human beings. We also validate biological meaning of modules identified by the proposed method.

Finally, we introduce random under-sampling techniques with application to X-ray computed tomography (CT). Under-sampling techniques are realized to be powerful tools of reducing the scale of problem especially for large data analysis. However, information loss seems to be un-avoidable which inspires different under-sampling strategies for preserving more useful information. Here we focus on under-sampling for the real-world CT reconstruction problem. The main motivation is to reduce the total radiation dose delivered to patient which has arisen significant clinical concern for CT imaging. We compare two popular regular CT under-sampling strategies with ray random under-sampling. The results support the conclusion that random under-sampling always outperforms regular ones especially for the high down-sampling ratio cases. Moreover, based on the random ray under-sampling strategy, we propose a novel scatter removal method which further improves performance of ray random under-sampling in CT reconstruction.

Table of Contents