Abstract

Feature selection (FS) aims to select a useful subset of features for classification while discarding irrelevant ones. Feature selection is challenging for low sample-size, high-dimensional data (e.g., neuroscience, gene expression datasets). The state-of-the-art FS techniques employ information-theoretic measures that are sensitive to the sample size resulting in spurious estimates of feature relevance, thereby deteriorating the classification and stability performance. In the first and second contributions of the thesis, we propose graph-based instance voting approaches for FS in small sample-size, high-dimensional data. Here, we represent each feature dimension as a graph whose nodes and edges correspond to the samples and their pairwise proximity, respectively. Graphs induced by the useful features will tend to exhibit a clustering (modular) structure of the nodes; we define instance votes as each node's contribution to graph modularity. For a given instance, its votes to the individual features reflect their usefulness for classification in the locality of the instance (local relevance). Next, we pose FS akin to the set-covering problem; we propose greedy heuristic strategies to select a relevant and non-redundant subset of the features such that they cover the instance space. The proposed FS algorithms yielded lower misclassification rate as compared to other considered FS algorithms on both synthetic and benchmark data. Through experiments on synthetic data, we demonstrate that the proposed FS algorithms can effectively suppress irrelevant features and yield reasonably stable feature subsets. Further, the proposed techniques outperformed other multivariate FS algorithms even for low training sample size and exhibited greater robustness to feature noise.

Next, we consider another related problem, kernel selection, that arises in multiple kernel learning (MKL). Kernel-based learning algorithms work upon pairwise similarity between the instances and can be used with both vector and non-vector (e.g. text, graphs) data inputs. Here, the similarity is computed using a specified kernel function or learned as data-dependent combination of multiple base

kernels (referred to as MKL). The kernel weights modulate each base kernel's influence in the overall representation. In the presence of a large number of base kernels, kernel selection (KS) is necessary to pre-select a subset of relevant and non-redundant kernel representations before combining them. While FS deals with individual/subsets of features in the feature space (i.e., Euclidean space), KS operates in the inner product space induced by the base kernels. In the third contribution of the thesis, we represent each base kernel as a signed weighted graph and show that the instance voting based greedy search developed for feature selection can also be used for kernel selection. Also, we propose a heuristic graph-based approach for multiple kernel combining; we define kernel weights in a localized manner, that allows them to vary across the input space. We demonstrate the effectiveness of the proposed approaches on synthetic and publicly available multi-view datasets.

Lastly, we demonstrate the application of the proposed algorithms to address the challenges in the classification of brain networks. We investigated functional connectivity networks (FCNs) of preschool children (36-59 months old) using continuous Magnetoencephalogram recordings acquired while they watched cartoon videos. Specifically, we considered the problem of classifying the children's performance on standardized assessment tests (Low vs High) from their FCNs. The classification of brain networks is a graph classification problem that can be addressed using either of feature-based or kernel-based frameworks: (a) In the former case, the number of node/edge-level features far exceeds the sample size. We employed the instance voting based FS techniques to select a stable subset of features for classification. Our results show that the selected features (that constitute the edges of the FCN) can reliably distinguish between the low and high scoring children. (b) The classification of brain networks using graph kernels may be dissociated into kernel selection (identifying the subset of useful nodes in the network) and multiple kernel combining (to obtain the overall pairwise graph similarity). Our results demonstrate that though global graph similarity measures may not be able to distinguish between classes, the local (node-wise) kernel measures are capable of capturing the differences in the functional brain network patterns with respect to the nodes under consideration.