早稲田大学大学院情報生産システム研究科

博 士 論 文 概 要

論 文 題 目

Study on Semi-Supervised Classification Based on Laplacian Kernel Machines Using Quasi-Linear Kernel

申 請 者

Yanni REN

情報生産システム工学専攻
ニューロコンピューティング研究

2021 年 12 月

Classification is one fundamental research topic in machine learning, which aims to recognize objects and separate them into classes. A classification model formulates the separation boundary between different classes, and generally, a nonlinear separation boundary is needed. Learning of a nonlinear classification model equals modeling a nonlinear separation boundary.

Traditionally, classification model learning has been studied in the supervised scheme where all the training data instances have accurate labels. However, labeled data is expensive in contrast to unlabeled data. Therefore, semi-supervised classification (SSC) has gained prominence, which leverages a large amount of unlabeled data in addition to a small amount of labeled data for training. Usually, intrinsic SSC methods are extensions of existing supervised methods to include unlabeled data in the objective function.

Laplacian kernel machines, namely, Laplacian Support Vector Machine (LapSVM) and Laplacian Regularized Least Square (LapRLS), are among the most result-promising semi-supervised classification methods. They are extensions of supervised kernel machines, Support Vector Machine (SVM), and Regularized Least Square (RLS) by adding a graph regularization in the objective function of model parameter optimization.

Kernel defines a linearly separable high-dimensional feature space, and a linear model in the feature space corresponds to a nonlinear model in the input space. The use of graph leverages unlabeled data by approximating data manifold, where data instance as nodes are sparsely connected by edges. The kernel is used again as edge weighting. Note that the kernel is used twice in Laplacian kernel machines, and its quality directly influences the performance of the classification model. General nonlinear kernels, such as radial basis function (RBF) kernels, implicitly define a general feature space. It is a black-box model from a modeling perspective, and prior knowledge cannot be used even if given.

In this dissertation, we are motivated to apply a two-step modeling method to model the nonlinear separation boundary using a set of linear models. The model parameters are estimated in two steps.

In the first step, the nonlinear parameters connecting or combining the linear models are estimated. Then the classification model is formulated as a regression form with a known regression vector and a parameter vector. The parameter vector contains all the linear parameters to be estimated in the second step. In the second step, the linear parameters of all the linear models are estimated globally. The classification model can be

further recast to a kernel form as an intermediate model. The kernel is defined as the inner product of the know regression vectors, namely, quasi-linear kernel. As a result, the quasi-linear kernel is composed in an interpretable way, and it contains prior knowledge.

Although the quasi-linear kernel has been studied in many tasks, exploiting it by leveraging a small amount of labeled data and a large amount of unlabeled data remains challenging. Therefore, we focus on its study in this dissertation to achieve accurate performance. We propose a series of semi-supervised classification algorithms based on Laplacian kernel machines through the construction of an intermediate model named quasi-linear kernel.

The dissertation contains the following five chapters as follows:

Chapter 1 first introduces the concepts mentioned above, such as nonlinear classification, semi-supervised classification, and Laplacian kernel machines. Then, we discuss the insufficiency of general kernels from the modeling perspective and introduce the two-step modeling method and the quasi-linear kernel. At last, we list challenges under the semi-supervised context, on which we will give corresponding solutions in the following chapters.

Chapter 2 proposes a Laplacian SVM based semi-supervised classifier using multi-local linear model. The semi-supervised classifier is constructed in two steps. In the first step, by applying a pseudo-labeling approach, the input space is divided into multiple local linearly separable partitions along the potential separation boundary. A multi-local linear model is then built by interpolating multiple local linear models assigned to the partitions. In the second step, the multi-local linear model is formulated as a linear regression form with a new regression vector containing the information of potential separation boundary. Then all the linear parameters are optimized globally by a LapSVM algorithm using a quasi-linear kernel function defined as the inner product of the new regression vectors. Furthermore, the quasi-linear kernel function and the pseudo labels are used to construct a label-guided graph. As a result, the potential separation boundary is detected, and its information is incorporated into a LapSVM in kernel and graph levels. Numerical experiments exhibit the effectiveness of the proposed method by showing better performance against general kernel LapSVM with a "win/tie/lose=7/1/0" on 8 real-world datasets under 10% labeled data.

Chapter 3 proposes a semi-supervised classifier based on piecewise linear model using gated linear network. The semi-supervised classifier

is constructed in two steps. In the first step, we design a label-guided autoencoder-based semi-supervised gating mechanism to generate binary sequences. By using a gated linear network, the binary sequences realize partitioning of a piecewise linear model indirectly. In the second step, the piecewise linear model is formulated as a linear regression form, and the linear parameters are then optimized globally by a LapRLS algorithm using a quasi-linear kernel function comprising the binary sequences. Moreover, the quasi-linear kernel function is used as a better similarity function for the graph construction. As a result, we estimate data manifold from both labeled and unlabeled data, and the data manifold is incorporated into both the kernel and the graph in LapRLS. The experimental results validate the effectiveness of the proposed method by showing a "win/tie/lose = 7/0/0" on 7 University of Cambridge Irvine (UCI) data sets compared to other SSC methods when 10% data is labeled.

Chapter 4 applies the proposed semi-supervised classifier based on piecewise linear model to parasite images, including a semi-supervised feature extractor based on deep CNN using contrastive learning.

First, for the deep CNN feature extractor, we introduce real-world images with similar and clear semantic information to enhance the structure at the representation level. In addition, we introduce variant appearance transformations to eliminate the texture at the representation level. Second, a gated linear network is adopted as the classifier to realize a piecewise linear separation boundary. The linear parameters are optimized globally by a LapSVM algorithm using a quasi-linear kernel function composed of the representations and the binary sequences generated from the learned feature extractor. In summary, the proposed semi-supervised method tackles the structure and texture challenges and achieves accurate parasite classification. The proposed method shows better performance than state-of-the-art SSC methods when only 1% of microscopic images are labeled. It reaches an accuracy of 95.10% in a generalized testing set.

Chapter 5 concludes the dissertation and provides future works. To conclude, this dissertation proposes a series of semi-supervised classification algorithms based on Laplacian kernel machines (Laplacian SVM, Laplacian RLS) through the construction of an intermediate model named quasi-linear kernel. In this way, we effectively leverage a small amount of labeled and a large amount of unlabeled data for training to achieve accurate performance on the testing set. Numerical simulation results on a wide range of benchmarks and real-world data sets demonstrate the effectiveness of the proposed semi-supervised classification algorithms.