Robust Twin Support Vector Machine for Noisy Data

A thesis submitted for the degree of Master of Science in Computer Engineering: Artificial Intelligence

by

Amir M. Mir

Thesis Advisor: Dr. Jalal A. Nasiri

ISLAMIC AZAD UNIVERSITY
North Tehran Branch

January, 2019

Summary

In the past decade, machine learning algorithms have been used for solving problems with complex patterns. Classification is the widely-used learning method, which solves complex learning problems such as face recognition, text classification, and medical diagnosis.

Support Vector Machine (SVM) is a powerful classification method. It is on the basis of Structural Risk Minimization (SRM) and VC-dimension. Due to the SRM principle, SVM has a great generalization ability. Therefore, it has been applied to a wide variety of applications. The main idea of the SVM is to find the optimal separating hyperplane with the largest margin between the two classes. To obtain such a hyperplane, SVM solves a convex Quadratic Programming Problem (QPP).

On the basis of standard SVM, scholars have proposed new classifiers in the past decade. For instance, Twin Support Vector Machine (TSVM) was proposed with the aim of classification using two non-parallel hyperplanes. Each of which fits the samples of its own class and is far from the samples of other class. Unlike SVM, TSVM solves two smaller-sized QPPs which makes its learning speed 4 times faster than that of SVM in theory. Although TSVM has better prediction accuracy and time complexity, it has several important drawbacks: (1) Its classification performance is sensitive to noise and outliers. (2) TSVM considers Empirical Risk Minimization (ERM). As a result, overfitting may occur. (3) The overall computational complexity of TSVM is also high for large-scale datasets.

With the aim of addressing the mentioned drawbacks of TSVM, this thesis proposes K-nearest Neighbor-based Least Squares Twin Support Vector Machine (KNN-LSTSVM), as well as a Regularized KNN-based Twin Support Vector Machine (RKNN-TSVM). The proposed KNN-LSTSVM has these advantages: (1) It uses the KNN graph to give weight to training samples that are located in highly-dense regions. Moreover, margin points of each class are determined. This improves the robustness of KNN-LSTSVM to noisy samples and outliers. (2) The least squares method was used to solve two systems of linear equations rather than solving two QPPs. This makes the learning speed of the proposed method faster than that of TSVM. Similarly, KNN-LSTSVM finds two non-parallel hyperplanes. However, each hyperplane fits the highly-dense samples of its own class and is far from the margin points of the other class. The experiment results on several synthetic and benchmark datasets indicate the effectiveness of the KNN-LSTSVM classifier in terms of classification accuracy and computational cost.

Inspired by TSVM and KNN-LSTSVM, the proposed RKNN-TSVM was designed, which has these advantages: (1) It gives weight to each training sample by considering the distance from its nearest neighbors. This further reduces the effect of noise and outliers on the output model. (2) To implement the SRM principle, a regularizer term was added to each objective function. Therefore, the generalization ability of RKNN-TSVM is improved. (3) To reduce the overall computational cost of the proposed method, location difference of multiple distances based k-nearest neighbors algorithm (LDMDBA) was employed to construct the KNN graph. The extensive experimental results on several synthetic and UCI datasets shows the efficiency of the RKNN-TSVM in terms of classification accuracy and computational time. Furthermore, the largest speedup in the proposed method reaches to 14 times.