Redefining the White-Box of k-Nearest Neighbor Support Vector Machine for Better Classification

Doreen Ying Ying Sim[1[006-082-593876]]

[1] Faculty of Cognitive Sciences and Human Development, University Malaysia Sarawak, Jalan Datuk Mohammad Musa, 94300 Kota Samarahan, Kuching, Malaysia
dsdoreenyy@gmail.com

Abstract. Distances and similarities among patterns of data points are computed by k-Nearest Neighbor methods after Principal Component Analysis is performed to the ten datasets. Weighted distances are then formulated, computed and adjusted synergistically with the Gaussian kernel width of Support Vector Machine. This is done by the proposed formulations of this research which is derived from the study on the relationships among the distances and similarities of patterns of data points as well as the kernel width of SVM. The kernel scale of Gaussian kernel width is customized and categorized by the proposed new approach. All these are known as the white-box algorithms which are to be redefined. The algorithm developed is to avoid and minimize the tradeoff and hinge loss problems of typical SVM classifications. After applying the proposed algorithms to the datasets mainly from UCI data repositories, it is shown to be more accurate in classification when compared with typical SVM classification without getting the Gaussian kernel width adjusted accordingly. Optimal kernel width from the customized kernel scale is input to the SVM classification after being computed by the proposed formulations. It is found that dimensionality reduction by PCA and distances among patterns computed by kNN and thereafter by the proposed formulations can optimally adjust the Gaussian kernel width of SVM so that classification accuracies can significantly be improved.

Keywords: k-Nearest Neighbor, Principal Component Analysis, Gaussian kernel width, Support Vector Machine, proposed formulations.

1 Introduction

1.1 Classical White-Box Algorithms of SVM and PCA

Support vector machine (SVM) is a stable and strong classifier [4]–[5], [9]–[12] and a supervised learning method for its classification and regression strength, while Principal Component Analysis (PCA) and k-Nearest Neighbor (kNN) are unsupervised methods. PCA is a type of dimensionality reduction techniques [1]–[3], [10]–[12] used to tune and monitor the white-box algorithms of SVM in this research. Both SVM and PCA are very popular machine learning algorithms [6]–[8], [10]–[12]. When data is visually and/or logically not easily separable, data might not be easily separated by any hard-margin, SVM usually uses a soft-margin to separate data.