Dimension Augmenting Vector Machine: A new General Classifier System for Large p Small n problem

Dipak K. Dey*, Samiran Ghosh, Yazhen Wang

Department of Statististics, Uuniversity of Connecticut, Storrs

Department of Mathematical Sciences, Indiana University-Purdue University, Indianapolis

Department of Statistics, University of Connecticut, Storrs

E-Mail: dipak.dey@uconn.edu

Abstract: Support vector machine (SVM) and other reproducing kernel Hilbert space (RKHS) based classifier systems are drawing much attention recently due to its robustness and generalization capability. All of these approaches construct classifier based on training sample in a high dimensional space by using all available dimensions. SVM achieves huge data compression by selecting only few observations lying in the boundary of the classifier function. However when the number of observations is not very large (small n) but the number of dimensions are very large (large p) then it is not necessary that all available dimensions are carrying equal information in the classification context. Selection of only useful fraction of available dimensions will result in huge data compression. In this paper we have come up with an algorithmic approach by means of which such an optimal set of dimensions could be selected. We have reversed and modified the solution proposed by Zhu and Hastie in the context of Import Vector Machine (IVM), to select an optimal sub model by using only few observations. For large p small n domain (e.g. Bioinformatics) our method compares different trans-dimensional model to come up with optimal set of dimensions to build the final classifier.