RotaSVM: A New Ensemble Classifier

Shib Sankar Bhowmick\textsuperscript{1,2,*}, Indrajit Saha\textsuperscript{1,*}, Luis Rato\textsuperscript{2}, and Debotosh Bhattacharjee\textsuperscript{1}

\textsuperscript{1} Department of Computer Science and Engineering, Jadavpur University, Kolkata 700032, West Bengal, India
\textsuperscript{2} Department of Informatics, University of Evora, Evora 7004-516, Portugal
shibsankar.ece@gmail.com, indra@icm.edu.pl, lmr@di.uevora.pt, debotosh@ieee.org

Abstract. In this paper, an ensemble classifier, namely RotaSVM, is proposed that uses recently developed rotational feature selection approach and Support Vector Machine classifier cohesively. The RotaSVM generates the number of predefined outputs of Support Vector Machines. For each Support Vector Machine, the training data is generated by splitting the feature set randomly into $S$ subsets. Subsequently, principal component analysis is used for each subset to create new feature sets and all the principal components are retained to preserve the variability information in the training data. Thereafter, such features are used to train a Support Vector Machine. During the testing phase of RotaSVM, first the rotation specific Support Vector Machines are used to test and then average posterior probability is computed to classify sample data. The effectiveness of the RotaSVM is demonstrated quantitatively by comparing it with other widely used ensemble based classifiers such as Bagging, AdaBoost, MultiBoost and Rotation Forest for 10 real-life data sets. Finally, a statistical test has been conducted to establish the superiority of the result produced by proposed RotaSVM.

Keywords: Principal component analysis, rotational feature selection, statistical test, support vector machine.

1 Introduction

Integration of classifiers nowadays is drawing much attention of the machine learning and pattern recognition communities and growing rapidly \cite{1,10}. In integrated classification techniques, an ensemble of classifiers is generated by giving similar or different permutated training data sets. Thereafter, class label of the test sample is assigned by either majority voting or averaging the output probabilities of the ensemble. Recent research shows that ensemble based classifiers, such as Bagging \cite{11}, AdaBoost \cite{12,13}, Random Forest \cite{14} and Rotation

* Both the authors contributed equally.
Forest\cite{15}, are used more often to increase the prediction accuracy of learning systems\cite{9,16,20}.

Among these ensemble classifiers, Rotation Forest\cite{15} performs much better than other ensemble methods. It uses rotational feature sets for decision tree classifiers. In Rotation Forest, rotational feature sets subsequently undergo Principal Component Analysis (PCA) to preserve the variability information of the training data. Here the main idea is to simultaneously increase diversity and individual accuracy within the decision tree classifiers. Diversity is achieved by using PCA, which is used to extract the principal components of rotational features for each classifier and accuracy is sought by keeping all principal components\cite{15,21}.

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser et al.\cite{22}. The basic idea of SVM is to find a hyperplane which separates the $d$-dimensional data perfectly into two classes. However, since classification data is often not linearly separable, SVM introduced the notion of a “kernel induced feature space” which embed the data into a higher dimensional feature space where the data is linearly separable. For this purpose, first the hyperplane is found, which separates the largest possible fraction of points such that points on the same side belong to the same class, while the distance of each class from the hyperplane is maximized.

As both Rotation forest and Support Vector Machine are successfully used in classification, thus their integration may achieve even higher prediction accuracy than either of them. Hence, in this paper, an ensemble classifier, named as RotaSVM, is proposed by integrating rotational feature selection scheme with SVM. The RotaSVM produces the number of predefined outputs of SVMs. For each SVM, the training data is generated by splitting the feature set randomly into $S$ subsets. Subsequently, principal component analysis is used for each subset to create new feature sets and all the principal components are retained to preserve the variability information of the training data. Thereafter, such features are used to train a Support Vector Machine. During the testing phase of RotaSVM, the sample data are the input to the rotation specific Support Vector Machines. Subsequently, it is classified by computing average posterior probability. The experimental studies were conducted with available 10 real-life data sets\cite{23}. The results show that RotaSVM can produce significantly lower prediction error more often than Rotation Forest and other ensemble based classifiers such as Bagging, AdaBoost and MultiBoost for all the data sets. Finally, $t$-test\cite{24} has been conducted to establish the statistical significance of the result produced by RotaSVM.

The rest of this paper is organized as follows: Section 2 briefly describes the Support Vector Machine classifier. The proposed RotaSVM is discussed in Section 3. Section 4 shows the empirical results. Finally, Section 5 concludes this paper with an additional note of future work.

\footnote{UCI repository\cite{23}.}