SWITCH: A Novel Approach to Ensemble Learning for Heterogeneous Data

Rong Jin[1], Huan Liu[2]

[1] Department of Computer Science and Engineering, Michigan State University
East Lansing, MI48824, U.S.A.
rongjin@cse.msu.edu
[2] Department of Computer Science and Engineering, Arizona State University
Tempe, AZ85287-8809, U.S.A.
hliu@cse.asu.edu

Abstract. The standard framework of machine learning problems assumes that the available data is independent and identically distributed (i.i.d.). However, in some applications such as image classification, the training data are often collected from multiple sources and heterogeneous. Ensemble learning is a proven effective approach to heterogeneous data, which uses multiple classification models to capture the diverse aspects of heterogeneous data. If an ensemble can learn the relationship between different portions of data and their corresponding models, the ensemble can selectively apply models to unseen data according to the learned relationship. We propose a novel approach to enable the learning of the relationships between data and models by creating a set of 'switches' that can route a testing instance to appropriate classification models in an ensemble. Our empirical study on both real-world data and benchmark data shows that the proposed approach to ensemble learning can achieve significant performance improvement for heterogeneous data.

1 Introduction

The standard framework of machine learning problems assumes that the available data is independent and identically distributed (i.i.d.), which usually results in a homogeneous distribution. However, in some applications such as image classification, the training data are often collected from multiple sources and thus exhibit a heterogeneous distribution. For heterogeneous data, a single classification model may not be sufficient to describe all the training data very well. One intuitive solution is to divide the heterogeneous training data into a set of homogeneous partitions and train a classification model over each homogeneous partition. To predict the class label for a testing instance, we can first examine which models are most likely to give a correct prediction for this instance and then apply those models to predict the class labels.

The idea of ensemble methods is to create multiple models from a single training dataset and combining them for classification. There have been many studies on this subject [1-3]. The well-known ensemble learning algorithms include Bagging [4], Gaussian mixture model (GMM) [5], and AdaBoost [6]. In this paper, we propose a novel ensemble learning approach that first partitions the heterogeneous data into homogeneous sections and then builds a classification model for each homogeneous section. Unlike most existing ensemble learning method where different models are combined linearly, the presented ensemble approach introduces a routing 'switch' for each classification model that automatically determine whether the classification model should be applied to input instances. With the switches, a different subset of classification models is invoked for each instance

2 SWITCH - A Novel Ensemble Approach

The new ensemble approach will be described in two parts: model generation and model combination.

Model Generation. Our approach toward model generation is to divide the training dataset into multiple homogeneous sections and create a classification model for each partition. One apparent approach for obtaining the homogeneous partitions is to apply some traditional clustering algorithm to group similar training data together. However, the drawback of this approach is that each partitioned section will only contain a small number of training examples and thus the resulting classification model can severely over-fit the partitioned data. To solve this problem, we combine multiple partitions together for training a single classification model. More specifically, in our experiments, we apply the EM clustering algorithm to divide the training data into 6 different partitions and a different classification model is trained for every two partitions. As a result, there are a total of 15 classification models and each one is trained over roughly 1/3 of the training data.

Model Combination. Let **x** be an instance, y be a class label, and $M = \{m_1, m_2, ..., m_n\}$ be the ensemble of n classification models. Our goal is to compute $P(y \mid \mathbf{x}, M)$. Let h_i be the hidden variable that indicates whether model m_i should be used for classifying the instance **x**. By assuming that the selection of one classification model is independent from the selection of another, likelihood $P(y \mid \mathbf{x}, M)$ can be simplified as the following expression:

$$P(y \mid \mathbf{x}, M) \approx \sum_{i=1}^{n} P(h_i = 1 \mid \mathbf{x}) P(y \mid \mathbf{x}, m_i) + Const \tag{1}$$

where models within the ensemble M are combined through another set of models $P(h_i = 1 \mid \mathbf{x})$. Details of derivation can be found [7] In our experiment, a model for estimating $P(h_i = 1 \mid \mathbf{x})$ can be learned as follows: For every classification model m_i in the ensemble, apply it to classify all training data and compare the predicted class labels to the true ones. For every training instance, if the predicted class label is the same as the true class label, mark it with a pseudo class '1'. Otherwise, a pseudo class '0' is assigned to the instance. Then, a 'switch', namely a classifier that is able to determine whether the corresponding model in the ensemble should be used to classify an instance, is trained over all the training instances with their pseudo class labels. The 'switch' will then be used to estimate the conditional probability $P(h_i = 1 \mid \mathbf{x})$. In the experiments below, a Naïve Bayes model is to estimate $P(h_i = 1 \mid \mathbf{x})$ due to its simplicity and reasonable good performance [8].

3 Experiments

Two heterogeneous datasets are used in this experiment: a dataset for indoor classification that contains 2500 examples represented by 190 features and a dataset for outdoor classification that contains 1403 examples represented 126 features. They are used to train image classifiers for identifying indoor and outdoor scenes. The experiments are performed with 5-fold cross validation and the average classification error rates are calculated. Table 1 summarizes the results for the new ensemble approach 'SWITCH' together with the baseline approach and two ensemble algorithms. For all methods, support vector machine is used as the basis model. We notice in Table 4 that compared to the baseline model, SWITCH is the only approach that consistently reduces the classification error significantly over the two datasets. It can be seen that Bagging also slightly outperforms the baseline model consistently over the two datasets. This result indicates that the new ensemble approach is effective for heterogeneous datasets. Italic numbers indicate they are smaller than the baseline error rates; italic and boldfaced numbers indicate the lowest error rates. In addition, more empirical studies of the proposed ensemble approach on both heterogeneous and homogeneous data can be found in [7].

	Outdoor	Indoor
SVM	0.238±0.032	0.463±0.036
AdaBoost	0.240±0.028	*0.442±0.044*
Bagging	*0.233±0.022*	*0.447±0.036*
SWITCH	***0.196±0.028***	***0.407±0.035***

Table 1: Averaged classification errors.

References

[1]. Dietterich, T.G., *An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization.* Machine Learning, 2000. **40**(2): p. 139-157.

[2]. Ali, K.M. and M.J. Pazzani, *Error Reduction through Learning Multiple Descriptions.* Machine Learning, 1996. **24**(3): p. 173-206.

[3]. Bauer, E. and R. Kohavi, *An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting and Variants.* Machine Learning, 1999. **36**: p. 105-139.

[4]. Breiman, L., *Bagging Predictors.* Machine Learning, 1996. **24**(2): p. 123-140.

[5]. Bishop, C.M., *Neural Networks for Pattern Recognition*. 1995, Oxford: Oxford University Press.

[6]. Schapire, R.E. and Y. Singer, *Improved Boosting Algorithms using Confidence-rated Predictions.* Machine Learning, 1999. **37**(3): p. 291-336.

[7]. Jin, R. and H. Liu. *SWITCH: A Novel Approach to Ensemble Learning for Heterogeneous Data.* MSU-CSE-04-24, Dept. of Computer Science and Engineering, Michigan State University, 2004

[8]. Domingos, P. and M.J. Pazzani. *Beyond independence: conditions for the optimality of the simple Bayesian classifier.* in Proceedings of the Thirteenth International Conference on Machine Learning. 1996.