A Review of Heterogeneous Ensemble Methods

Sam Reid
Department of Computer Science
University of Colorado at Boulder

February 15, 2007

Draft Version

Abstract

Several ensemble methods have been proposed that can accommodate differing base model types. This document reviews the recent literature, and for each method, we identify (1) main contributions, (2) theoretical motivation, (3) empirical results and (4) relationships to other techniques.

1 Introduction

Ensemble methods combine the predictions of multiple base models[1] [4]. Base models can be created by resampling, manipulation or randomization of the training data, learning algorithms or learning parameters. Theoretical [7] and empirical [11][5] results have demonstrated the advantage of ensembles over single models. Homogeneous ensemble methods use the same base learner on different distributions of the training set, e.g. bagging and boosting. Heterogeneous ensemble methods incorporate different model types into the library of models, the idea being that different base model types can be both accurate and diverse. For heterogeneous ensembles, ensemble creation is usually a 2-phase process (sometimes called overproduce & select): many different base models are generated by running different learning algorithms on the training data, then the generated models are combined to form the

[1] or 'hypotheses', 'predictors', 'generalizers'

ensemble. Several methods have been proposed for constructing heterogeneous ensembles, but comparisons are difficult to make since these methods are typically evaluated using different model libraries and different datasets. Heterogeneous ensemble methods may be compared by looking at performance metrics as a function of ensemble size, where each ensemble method is given the same model library for selection. It could also be worthwhile to analyze various diversity/accuracy or bias/variance measures as a function of ensemble size for different methods in order to understand these tradeoffs. A thorough comparison of the empirical results of different heterogeneous ensemble methods along with their theoretical relationships could lead to an (1) improved understanding of existing ensemble methods, (2) an understanding of what theories at work in existing ensemble methods seem to be most important, (3) development of improved ensemble methods.

Contents

2 Basic & Generalized Ensemble Methods

Perrone & Cooper [12] proposed two methods for constructing ensembles. Though their empiricial results are applied only to ensembles of neural networks, their framework applies to heterogeneous classifiers. The Basic Ensemble Method (BEM) simply averages the outputs of all base models, $f_{BEM}(x) = 1/N \sum_{i=1}^{N} f_i(x)$. The success of the BEM hinges on two assumptions: base models are independent and the error of each base model has zero mean. These two assumptions are often violated in practice. The Generalized Ensemble Method (GEM) discards base models that are too closely correlated (as measured by similarity of predictions), and uses a closed form expression to determine weights α_i for each model. The model is $f_{GEM}(x) = \sum_{i=1}^{N} \alpha_i f_i(x)$, where $\sum_{i=1}^{N} \alpha_i = 1$. The weights α_i are computed as $\alpha i = \sum_j C_{ij}^{-1} / \sum_k \sum_j C_{kj}^{-1}$, where C_{ij} is the symmetric correlation matrix $C_{ij} = E[m_i(x)m_j(x)]$ and $m_i(x)$ is the misfit function $m_i(x) = f(x) - f_i(x)$. The success of GEM depends on the invertibility and reliable estimation of C_{ij}. If these two assumptions hold, GEM can be shown to provide the best estimate of $f(x)$ under the mean square error loss function[12].

Empirical Studies GEM was studied on the NIST OCR database, and showed improvements over the single-model predictor.

Comments This method should be compared to many other models by evaluation on a variety of UCI datasets.

3 Stacking

Stacking, or stacked generalization, trains a meta-level model on the inputs combined with the outputs of the ensemble members[15]. The meta-level model is thus able to incorporate information from the base models into its decision, based on the location of the test data point. [5] describes an extension to the original stacking framework that improves performance over the

single best model (chosen by cross validation), when using a heterogeneous model library.

Empirical Studies Experiments are run with heterogeneous base model types and over 30 standard (mostly UCI?) datasets, and comparisons are made to several incarnations of the stacking algorithm.

Comments

4 Greedy Search

Ensemble Selection from a Library of Models [2][1] uses stepwise forward greedy search to construct the ensemble, with optional preprocessing steps to prune inaccurate model library members and to initialize the ensemble with the N best model library members. CV-Committees are shown to be important in improving the accuracy of the ensemble. Bagged model selection is used to reduce the probability of overfitting to the validation set.

Empirical Studies Base models consist of SVM, kNN, DT, BAG-DT, BST-DT, BST-STMP with many different learning parameters. Experiments are also run with respect to a variety of different performance metrics, for 7 problems, including binary UCI repository problems and a SLAC dataset. Results are compared to a variant of Bayesian model averaging.

Comments It could be rewarding to analyze how this greedy search is managing the accuracy/diversity and bias/variance tradeoffs.

5 Ensemble Pruning via Semidefinite Programming

Zhang et al. [17] describes a semi-definite programming approach to convex ensemble pruning. The technique is proposed as a pruning method, but this is equivalent to selecting an effective subset of models from a model library as in the overproduce & select framework. The subset selection is formulated as a quadratic integer programming problem (which is NP-hard)

and transformed into a convex semidefinite programming problem, which is solvable in polynomial time.

Empirical Studies Ensemble Pruning via Semidefinite Programming is evaluated on 16 binary UCI datasets (some of which have been converted to binary). The model library is a collection of 100 decision trees generated from (possibly multiple) runs of Adaboost. The group reports occasional improved performance over the total model library as a weighted ensemble.

Comments This technique should be evaluated on a heterogeneous model library. Are there tractable extensions of this algorithm that can be used on very large model libraries?

6 Statistical Ensemble Method

Yáñez et al. propose using resampling & statistical techniques for estimating base model generalization error and multiple comparison techniques for selecting models from the model library[6]. A single evaluation on a cross-validation set may yield an inaccurate measure of the base model's performance over the complete probability distribution of inputs. Models are added to the ensemble if they have similar accuracies (with statistical significance $\alpha = 5\%$), and combined by unweighted voting. Related work investigates more flexible distributions over the weights [Yáñez thesis]. This work is very similar to [13].

Empirical Studies The Statistical Ensemble Method is evaluated on 6 datasets from UCI, Donoho-Johnstone and ELENA Projects. Base models are RBF networks with different numbers of kernels.

Comments This method should be studied in the presence of heterogeneous models.

7 Effective Voting of Heterogeneous Classifiers

Tsoumakas et al. computes a significance index for each model in the library, and proposes 3 strategies for selecting models based on the significance results [13]. This work is closely related to the work in[6], with differing weight allocation during ensemble construction.

Empirical Studies This approach is evaluated on 40 UCI datasets, with 10 base-level algorithms. Default parameters are used for each base-level algorithm. The approach is compared to stacking.

Comments How does this approach perform in the presence of a large model library in which base algorithms are run multiple times with different learning parameters?

8 Competence Analysis

Jankowski & Grabczewski describe a framework in which global and local 'competence' is used to construct ensembles based on base-model accuracy in different regions of the input space[8]. They also describe configurations based on member type (single member or CV-committee), committee type (weight distribution), global competence and local competence.

Empirical Studies 15 different committee configurations are studied, with kNN, DT, Naive Bayes and SVM base models. Tests were performed on 17 UCI repository datasets.

Comments This technique should be compared to other techniques (e.g. stacking).

9 Clustering

Li et al. cluster outputs of model library members to select diverse models for participation in the ensemble[16]. The ensemble's predictions are made by voting.

Empirical Studies Experiments are performed over 10 UCI repository datasets.

Comments This technique needs to be compared to competing techniques.

10 Bayesian Sum-Of-Trees

Chipman et al. introduced a Bayesian framework to produce a sum-of-trees ensemble of weak models[3]. Monte Carlo sampling is used to sample from the distribution defined by the prior and likelihood.

Empirical Studies This technique is compared to boosting, neural nets, linear regression and random forests on Boston Housing data and Friedman's synthetic data.

Comments Can these techniques be applied to heterogeneous model types?

11 Bayesian Classifier Combination

Ghahramani and Kim describe a Bayesian framework for correlated classifiers and describe the inherent flaws with Bayesian Model Averaging (BMA), namely that BMA assumes classifier independence, mutual & exhaustive coverage, existence of the likelihood distributions and conditioning on the same dataset.

Empirical Studies Experiments are performed on Satellite and DNA datasets from Statlog and the UCI digit dataset. Comparisons are made between homogeneous and heterogeneous ensembles.

Comments The authors discuss extending this work to handle probabilistic outputs from the base models. This sounds appropriate, and may benefit from calibration.

12 Using Unlabeled Data for Ambiguity

Krogh & Vedelsby define a term 'ambiguity' as a proxy for base model diversity[9]. Unlabeled data is used to estimate the base model ambiguity, which is subsequently used to determine the weight distribution for ensemble members. This work is related to a resampling ensemble technique called DECORATE[10] in which extra-sample data is synthesized in order to ensure base model ambiguity.

Empirical Studies This technique is only studied for neural network base model types.

Comments

13 Bias/Variance Decomposition

Wichard et al. looks at the bias/variance decomposition for heterogeneous ensembles in theory and with respect to a set of empirical experiments[14].

Empirical Studies The Friedman synthetic data and Abalone datasets are analyzed. Monte carlo techniques are used to estimate the bias, variance and generalization error. Ensembles of ANN, RBF, kNN and polynomial regression models are compared to an ensemble containing all of these base models.

Comments

14 Conclusions

Proposed heterogeneous ensemble techniques have been rarely been compared to one another. A theoretical and empirical comparison of proposed heterogeneous ensemble methods could lead to an improved understanding of ensemble techniques and promising avenues for future research.

References

[1] R. Caruana, A. Munson, and A. Niculescu-Mizil. Getting the most out of ensemble selection. In *ICDM*, December 2006. Full-length version available as Cornell Technical Report 2006-2045.

[2] R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes. Ensemble selection from libraries of models. In *Proc. 21st International Conference on Machine Learning (ICML'04)*, 2004.

[3] H. Chipman, E. George, and R. McCulloch. Bayesian ensemble learning. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

[4] T. G. Dietterich. Ensemble methods in machine learning. *Lecture Notes in Computer Science*, 1857:1–15, 2000.

[5] S. Dzeroski and B. Zenko. Is combining classifiers with stacking better than selecting the best one? *Mach. Learn.*, 54(3):255–273, 2004.

[6] A. Y. Escolano, P. G. Riaño, J. P. Junquera, and E. G. Vázquez. Statistical ensemble method (sem): A new meta-machine learning approach based on statistical techniques. In *IWANN*, pages 192–199, 2005.

[7] L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.

[8] N. Jankowski and K. Grabczewski. Heterogenous committees with competence analysis. In *HIS '05: Proceedings of the Fifth International Conference on Hybrid Intelligent Systems*, pages 417–424, Washington, DC, USA, 2005. IEEE Computer Society.

[9] A. Krogh and J. Vedelsby. Neural network ensembles, cross validation, and active learning. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7, pages 231–238. The MIT Press, 1995.

[10] P. Melville. *Creating Diverse Ensemble Classifiers to Reduce Supervision*. PhD thesis, University of Texas at Austin, 2005.

[11] D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[12] M. P. Perrone and L. N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In R. J. Mammone, editor, *Neural Networks for Speech and Image Processing*, pages 126–142. Chapman-Hall, 1993.

[13] G. Tsoumakas, I. Katakis, and I. Vlahavas. Effective voting of heterogeneous classifiers.

[14] J. Wichard and C. Merkwirth. Building ensembles with heterogeneous models, 7th course of the international school on neural nets iiass, 22-28 sep. 2002, salerno, italy. citeseer.ist.psu.edu/wichard03building.html.

[15] D. H. Wolpert. Stacked generalization. Technical Report LA-UR-90-3460, Los Alamos, NM, 1990.

[16] K. L. Ya-Min Li, Li-Juan Cui. Creating diversity in ensembles using clustering method from libraries of models. In *Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian*, 2006.

[17] Y. Z. Yi-Zhang. Journal of machine learning research 7 (2006) 1315–1338 submitted 8/05; revised 4/06; published 7/06 ensemble pruning via semi-definite programming.