SAND2010-0759C

Mixtures of Local Experts for Supervised Classification on Heterogeneous Data

Submitted for Blind Review to AAAI 2010, Paper ID #855
Topics Areas – Machine Learning Classification, Ensemble Methods, and Data Mining

Abstract

Data partitioning is a common approach to processing data sets too large to fit into the memory of the available processors. Partitioned data may, however, be inhomogeneous in its class or feature statistics. In such instances, a training data set associated with a particular partition will therefore likely not be representative of the general population, with the resulting local models generalizing poorly in some regions of feature space. In response, we present an ensemble approach which embraces the model variation induced by the various inhomogeneous partitions, combining those models in a way that makes use of each model's applicability to the specific test instance in question. We use the Mixture of Experts model as our framework for combining experts conditioned on the input, where normalized Gaussian density models learned at training time provide a soft decomposition of the input space. We evaluate our technique in a statistically significant study across multiple datasets, and show that our proposed local experts method statistically significantly outperforms non-local baselines.

1. Introduction

Data partitioning is a common approach to processing data sets too large to fit into the memory of the available processors; "divide and conquer" is often an effective mechanism for parallelism. One complication, however, is that the partitioned data may be inhomogeneous in its class or feature statistics. This occurs, for instance, in temporal data sets subject to concept drift, or in multi-processor simulations of physical phenomena where the partitioning was chosen for computational load-balancing, not statistical analysis.

When mining such datasets, a training data set associated with a particular partition will likely not be representative of the general population as a whole. In particular, the resulting local models may not be applicable to, and may not generalize well to, other regions of feature space because they have not been exposed to training instances from other partitions.

In response, we present an ensemble approach which embraces the model variation induced by the various inhomogeneous partitions. We build models on each partition and subsequently combine their outputs to classify new data. This combination must be done with respect to each model's applicability to the test instances; otherwise, models may be applied to test instances that come from distributions different from what the model was trained on. In particular, a *local* method that combines a classifier's probabilistic output with some measure of that classifier's *applicability* at some test instance $\mathbf{x}$ is needed.

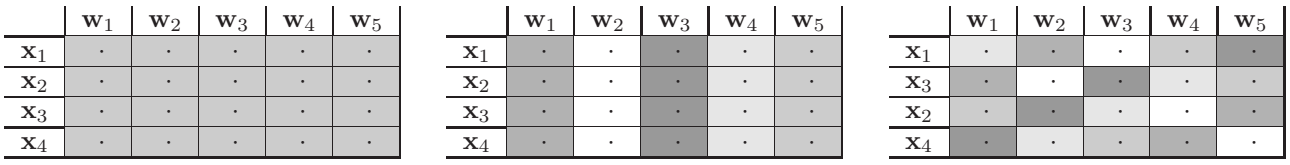
When conditioned on the input, such an applicability measure can be thought of as a *local accuracy estimate* (Woods, Kegelmeyer, and Bowyer 1997; Cevikalp and Polikar 2008). In contrast to such local methods, it is also possible to incorporate a measure of general prior confidence in a classifier, independent of the input; this globally weighted approach is considered as a baseline in the study done for this paper. Fig. 1 illustrates the key differences in the resulting weights assigned to the experts, as a function of test point, for both global and local approaches.

One way to derive a local accuracy estimate of an expert for a test instance $\mathbf{x}$ is to estimate the similarity of $\mathbf{x}$ to data on which the classifier was trained. While maintaining all the training data used to create a model is generally not feasible, estimating the distribution of training data with density models provides an efficient mechanism that can later be evaluated to provide estimates of model applicability. In this work, class-conditional density models, in the form of normalized Gaussians learned on training data, are used.

Early work in this area focused primarily on regression scenarios (Sato and Ishii 2000; Moody and Darken 1989), while the use of normalized Gaussian networks in non-stationary environments has also been previously investigated (Ramamurti and Ghosh 1999). Our application of these methods to cope with inhomogeneous data scenarios is novel, as is our particular instantiation of the Mixture of Experts model for this task.

Examples of Inhomogeneous Partitions

One spatially oriented example of inhomogeneous data partitions comes from the United States Department of Energy's Advanced Simulation and Computing (ASC) program (Kusnezov 2004), wherein a supercomputer simulates, for instance, the structural properties of a safety container. These simulations are very fine-grained and high-fidelity, and so require that the resulting simulation data, terabytes to petabytes in size, be partitioned and distributed across separate disks, to facilitate parallel computation. Since these partitions are defined *spatially*, they tend to be very inhomogeneous in their content, as each partition contains only a small part of a complex assembly.



(a) Unweighted Average (b) Weighted Average (c) Local Experts

Figure 1: Notional illustration of weight matrix $\mathbf{W}$ for Unweighted, Weighted, and Mixture of Experts (MoE) approaches to combining expert predictions. (Compare with Fig. 3, taken from experimental results.) Darker intensity indicates heavier weight in the final expert combination. The Local MoE method is able to weight the experts to varying degrees as a function of the test point, while the global Weighted Average approach cannot.

A second example relates to distributed mining of very large scale search query data, e.g., data stored by popular search engines. Such data is not only temporally inhomogeneous because search trends and user behaviors can change over time, but is also spatially inhomogeneous in that log data are stored in a distributed fashion. These concepts also apply more generally to parallel distributed database systems which often contain inhomogeneous data partitions.

Contribution and Novelty

The contributions of this paper are two-fold. First, we motivate, propose, and formulate a Mixture of Experts approach for combining predictions from multiple models to cope with inhomogeneous partitions; the use of local experts in such scenarios is novel. Second, we show in a statistically significant empirical study that the proposed Mixture of Experts method performs better on inhomogeneous partitions than several baseline alternative algorithms, and performs as well as a single model trained on all the data.

2. Mixture of Experts Approach

The Mixture of Experts (MoE) model (Jacobs et al. 1991; Jordan and Jacobs 1994; Jacobs 1995) is fundamentally a conditional mixture model in which the mixing coefficients, like the expert response, are functions of the input. Essentially, it is a mechanism for combining expert predictions, represented as component densities in the MoE model.

Background

Use of the MoE model is motivated in data mining scenarios involving inhomogeneous partitions, because experts trained on individual partitions will only be applicable (and appropriate for evaluation) on a subset of test instances whose distribution is unknown, i.e. that subset of instances which most closely resembles the instances on which the particular local expert was trained.

While some models do allow for combining multiple experts, and some even allow the specification of *a priori* belief in a particular expert, such *non-local* or global approaches cannot vary the weights assigned to a given classifier as a function of the input. Hence, with global weights, there is no way to disregard a particular model response for an input on which it would tend to misclassify (Fig. 1).

The local weights in the Mixture of Experts model are implemented by the mixing coefficients, which effect a decomposition of input space. In its most rigorous form, a *hard decomposition* can be extracted, with the resulting *gating network* identifying one single expert from some ensemble of such experts whose sole prediction is used as the final prediction at a particular test point $\mathbf{x}$. Generalizing, the model also permits the use of a *soft decomposition*, in which the outputs from *multiple* experts are considered at $\mathbf{x}$. Such a mixture can be thought of as a weighted mean of the expert responses conditioned on the input, as visualized in Fig. 1c. Note that the MoE model generalizes to an arbitrary number of classes.

The MoE model itself does not prescribe how to determine the mixing coefficients. We adopt prior work in this area (Ramamurti and Ghosh 1999; Procopio et al. 2009) that uses normalized Gaussians to determine a soft decomposition of input space. This specific approach is not novel; in particular, Sato and Ishii in (Sato and Ishii 2000) use the Normalized Gaussian Network (Moody and Darken 1989), or *NGnet*, as the basis for their proposed on-line EM algorithm, used in turn to fit model parameters. Our approach for determining the mixing coefficients is similar, although our end task is classification, not regression, and the data scenarios are considerably different. Note that we do not explicitly fit the model using EM as outlined in (Jordan and Jacobs 1994); rather, we directly derive mixing coefficients from learned density models.

MoE Model Specification

The Mixture of Experts model on which our approach is based is a type of *conditional mixture model* in which the mixing coefficients are functions of the input:

$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) p_k(\mathbf{t}|\mathbf{x}) \ . \tag{1}$$

Here, the individual component densities $p_k(\mathbf{t}|\mathbf{x})$ are the *experts*, the mixing coefficients $\pi_k(\mathbf{x})$ are known as *gating* functions, and $\mathbf{t}$ is the resulting vector of probability mass assigned to the class targets (Bishop 2006).

Mixing Coefficients: Density Models The mixing coefficients in this technique are determined by class-conditional Gaussian density models fit to training data when training the expert. When training expert k, a single multivariate Gaussian model $\mathcal{G}_{k,c}$ is learned for each class c of training data from the current partition. Because only a single Gaussian is used to estimate this density, the density model is determined directly from the sample mean and covariance of that data; use of the EM algorithm is not required. This is speedy, and as we show in this paper, effective. However, it also possibly underfits complicated cluster structures; see the "Future Work" section for our thoughts on relaxing this "single Gaussian per class" assumption.

During evaluation, for each test point $\mathbf{x}$, the mixing coefficients for model k are determined from the response of the density model $\mathcal{G}_{k,c}$ at $\mathbf{x}$:

$$\mathcal{G}_{k,c}(\mathbf{x}|\boldsymbol{\mu}_{k,c}, \boldsymbol{\Sigma}_{k,c}) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}'_{k,c}\right|^{1/2}} \times$$

$$\times \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{k,c})^T(\boldsymbol{\Sigma}'_{k,c})^{-1}(\mathbf{x}-\boldsymbol{\mu}_{k,c})\right], \quad (2)$$

where $\mathbf{x}$ is a d-dimensional feature vector, $\boldsymbol{\mu}_{k,c}$ is a d-dimensional mean vector, $\boldsymbol{\Sigma}'$ is a scaled $d \times d$ covariance matrix, and $|\boldsymbol{\Sigma}'|$ denotes the determinant of $\boldsymbol{\Sigma}'$. $\boldsymbol{\mu}_{k,c}$ and $\boldsymbol{\Sigma}_{k,c}$ are the sample mean and covariance, respectively, of the training data used to fit $\mathcal{G}_{k,c}$. The remaining details of the model are found in (Procopio et al. 2009).

MoE Model Implementation

The proposed Mixture of Experts approach can be efficiently implemented as a wrapper around both the training and evaluation portions of some existing classifier system. All that is required is that base learners yield predictions as a PDF over the C possible classes. Such probabilistic output is naturally obtained for many classifiers (e.g., Naïve Bayes), while other classifiers, such as the Support Vector Machine (Vapnik 1995), require special scaling.

3. Experimental Approach

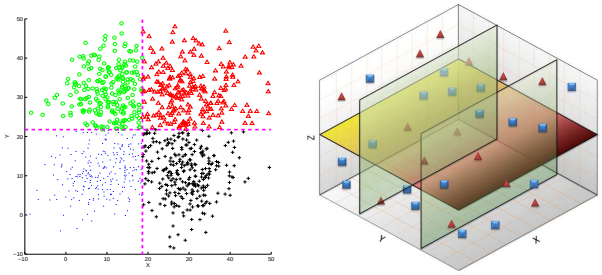
The primary research hypothesis of this study is that the local Mixture of Experts approach to combining expert output will outperform non-local techniques such as the Unweighted and Weighted Average baselines, while approaching the performance of that of single-model and ensemble-based "Sages" trained on all of the data.

We conduct experiments under two scenarios. The first scenario will examine *heterogeneous partitions*, using the partitioning scheme described below to create partitions with differing data distributions. In contrast, the second experimental scenario will examine *homogeneous partitions*, created by randomly sampling from the general population, i.e., all of the data. Though we are primarily interested in measuring the improved accuracy of our methods when applied to heterogeneous data, we examine homogeneous data as well, to assess whether the proposed local MoE method does worse, compared to our baselines, in that context.

Partitioning for Simulated Heterogeneity

We wish to evaluate our method with data that will both permit determination of statistical significance and also allow replication by others. It would be straightforward to generate simulated data, but it is often difficult to generalize from performance on simulated data to real world data. More attractive would be a method for stratifying real data that would yield heterogeneous partitions of data.

Principled partitioning methods, also referred to as binning or discretization methods, are known and are described in the literature (Han 2005). *Equal-width* (distance-based) partitioning divides a feature into N intervals of equal size; the resulting partitions will contain varying numbers of data instances (samples). In contrast, *equal-depth* (frequency-based) partitioning divides a feature range into N intervals, which may be of varying size. The resulting partitions will each contain approximately the same number of samples.



(a) Example 2D Partitioning (b) Example 3D Partitioning

Figure 2: Examples of feature space partitioning. Split locations are determined by dividing the range of the specified attribute into even-sized partitions, without regard to class labels or the distribution of the data.

In this research, we used an equal-width partitioning scheme, where we choose a subset of $F = 3$ features in each data set, splitting each feature once into $N = 2$ equal-sized partitions. This generates $2^3 = 8$ partitions which are likely heterogeneous in feature value distribution. To avoid bias, we choose the set of partition features $\mathcal{F}$ in an exhaustive, one-time search procedure that minimizes the variance of the number of examples falling in each partition. A future study will also investigate performance using the equal-depth partitioning approach.

Effective Homogeneity

To effectively assess performance on inhomogeneously partitioned data, we first need to be able to quantify the degree of inhomogeneity. We propose a new approach to evaluate the *effective homogeneity* of the partitions, which is not explicitly tied to the actual data, but rather to the performance of local experts across all partitions. In particular, our method measures homogeneity indirectly by considering the performance degradation of non-native local experts, i.e., experts trained on different partitions, versus the native expert. The essence of the idea is that the greater the difference in the distribution of data across partitions, the larger the performance differential will be for a local versus non-local expert applied to a given partition.

The homogeneity H is determined by linear interpolation onto the interval $[0, 1]$ by bounding by the worst-case score (random performance, R) and the best-case native score, $\mathbf{n}_i$). Here, "native" refers to an expert associated with the same data partition on which it was trained. For K experts,

$$H(\mathbf{o}, \mathbf{n}, R, K) = \frac{1}{K}\sum_{k=1}^{K}\frac{(\mathbf{o}_k - R)}{(\mathbf{n}_k - R)}, \quad (3)$$

where $\mathbf{n}$ is the vector of native scores for the partitions; $\mathbf{o}$ is the vector containing the mean of the scores at each partition of all non-native experts evaluated on the data from that partition; and R as the worst-case (random) performance of the the classifier ($R = 1/C$ where C is the number of classes).

The proposed homogeneity measure is comparable in a relative sense, and also has meaning in an absolute sense.

Further, our approach generalizes to arbitrary numbers of classes and partitions. Homogeneity scores are reported as continuous numeric values between 0 and 1; 0 is maximally heterogeneous, and 1 is fully homogeneous. Table 1 includes the effective homogeneity scores for our datasets after partitioning by the method described in the previous section.

Evaluation Method

For both experimental scenarios, 8 disjoint partitions in feature space are created, according to the equal-width partitioning scheme presented earlier in Sec. 3. A single expert is trained on each partition, resulting in 8 total experts. (An expert comprises both a base classifier model as well as a set of class-conditional density models that determine classifier applicability; see Eq. 1).

The classifier portion of each expert is evaluated independently on holdout data. The response from each expert is then combined by one of three methods: Unweighted Average, Weighted Average, or Local Mixture of Experts (see Fig. 1). The class receiving the most weight in the combined output is the final expert prediction. For each test point in the holdout data, this final prediction is compared with ground truth; over the entire test set, classifier accuracy (proportion of correct predictions) is reported.

The base classifier throughout this study is fixed as the Support Vector Machine (SVM) (Vapnik 1995) using the Radial Basis Function (RBF) kernel. Associated learning parameters for RBF-SVM (cost parameter c and RBF gamma g) are optimized during training using 10-fold cross-validation. LIBSVM v2.89 (Chang and Lin 2001) is used as the SVM implementation.

We conduct an empirical evaluation over five datasets from the UCI data repository, comparing the performance of the proposed approach versus four baseline algorithms; the datasets are summarized in Table 1. Stratified 5×2 cross-validation is used, in which a model is trained on one split of the data and tested on the second split; thus there are 10 randomized experiments in total.

Statistical Evaluation

The mean and standard deviation for the resulting set of 10 scores are reported in Table 2. The individual scores from the cross validation folds form the statistical basis for comparing classifiers using the Wilcoxon Signed-Ranks Test, a non-parametric analog to the paired t-test. All statistical tests are conducted at the 95% confidence level ($\alpha = 0.05$).

Algorithms

We compare the proposed local Mixture of Experts method's performance to that of four baseline algorithms: the Unweighted and Weighted Average methods discussed below, the trivial classifier that predicts the majority class, and two approaches referred to as the "Sages."

The *single-model* Sage is trained on all of the data from all of the partitions, and is therefore not disadvantaged by seeing only local partitions of data. The *multi-model* Sage also sees all the data, but uses Bootstrap Aggregating, or Bagging (Breiman 1996) to generate and vote over multiple models. For the Bagging Sage, we used $b = 8$ bags (bootstrap samples), in order to match the 8 experts created in the MoE approach (one expert for each partition; see Sec. 3.). Each bootstrap sample contained the same number of data instances as the training data set.

Table 1: UCI Datasets Used in the Evaluation

Dataset	Instances	Features	Classes	Homogeneity
adult	48842	14	2	0.83
krk	28056	6	18	0.23
letter	20000	16	26	0.38
nursery	12960	8	5	0.54
pendigits	10992	16	10	0.44

Datasets

The five datasets used in the evaluation comprise both binary and multiclass scenarios, and are associated with varying number of features, classes, and degree of class imbalance (skew). Datasets with a larger number of instances were preferred ($N \geq 10000$) in order to be more representative of larger-scale data mining scenarios. A summary of the datasets used in the evaluation is given above in Table 1.

4. Experimental Results and Discussion

Heterogeneous Partitions Scenario

The middle section of Table 2 presents the experimental results from the heterogeneous partitions scenario. In this scenario, models are trained on disjoint subsets of feature data, which were sampled in such a way as to have different distributions from the target population.

Performance of MoE Method When applied to heterogeneous partitions, the MoE method statistically significantly outperformed the non-local methods. The performance of the MoE approach also statistically significantly exceeded that of the two Sages on two datasets (*krk* and *nursery*). This is an important result; while all approaches saw all the same data, the MoE ensemble of local classifiers, constrained to training on disjoint, inhomogeneous subsets of data, outperformed both a single model and a bagged ensemble of models, both of which were able to see all the data at once. Moreover, on the remaining datasets, MoE approached the performance of the top-performing model to within 1.5%.

The power of the MoE method is illustrated by the weight matrix in Fig. 3b, taken directly from the experimental results on the *adult* dataset; this is a real data version of the notional example in Fig. 1c. Crucially, the MoE method is able to completely ignore the predictions of inappropriate models, i.e., models trained on partitions in which data do not look like the test point under consideration.

As Fig. 3b shows, most of the weight for a given test point is spread over one or two most applicable models. In comparing this matrix versus the Weighted Average weight matrix approach shown in Fig. 3a, we conclude that the performance benefit gained from the MoE approach is likely due to (a) exclusion of inapplicable models' predictions in the final vote by assigning low weight, and (b) appropriate consideration of multiple applicable models, beyond just the one trained on the partition that the current test point came from.

Table 2: Experimental results (classification accuracy, %) for baselines (left section), heterogeneous partitions scenario (middle section), and homogeneous partitions scenario (right section). **Boldface** indicates the highest accuracy within each section; asterisk (*) indicates highest accuracy overall.

	BASELINES			HETEROGENEOUS PARTITIONS			HOMOGENEOUS PARTITIONS		
DATASET	Predict All Major. Class	Single Model Sage	Bagging Sage	Unweighted Average	Weighted Average	Local MoE	Unweighted Average	Weighted Average	Local MoE
adult	76.07 ±.07	84.53 ±.34	***84.64 ±0.25**	76.67 ± .28	76.59 ± .27	**84.54 ±.20**	84.10 ±.23	84.10 ±.23	**84.14 ±.23**
krk	16.23 ±.27	56.45 ±.92	**56.61 ±0.44**	35.27 ± .79	33.96 ± .79	***58.49 ±.23**	48.05 ±.44	48.03 ±.42	**48.22 ±.33**
letter	4.22 ±.08	94.18 ±.87	***94.42 ±0.68**	87.10 ±1.08	87.03 ± .97	**92.95 ±.79**	90.02 ±.93	90.02 ±.83	**91.16 ±.89**
nursery	33.40 ±.45	**98.89 ±.27**	98.78 ±0.12	94.50 ± .45	94.25 ± .59	***99.34 ±.17**	96.89 ±.24	96.88 ±.24	**97.01 ±.21**
pendigits	10.74 ±.13	***99.53 ±.06**	99.50 ±0.10	86.68 ±2.17	86.72 ±2.61	**99.20 ±.09**	99.09 ±.15	99.09 ±.15	**99.28 ±.07**

Performance of Global Approaches Another important result was that for this heterogeneous partitions scenario, the Unweighted and Weighted Average global approaches—unable to derive and act on local applicability estimates—performed statistically significantly worse than the Sage and MoE methods. We conclude that the performance differential is due to MoE's locally aware combination scheme (i.e., the expert mixture). In particular, Gaussian density models are effective for determining correct mixing coefficients and informing where models are applicable.

Improved Performance versus Bagging Sage Although it is well known that use of ensemble methods (multiple classifiers) can yield improved classification performance, many of these results center around Bagging (Bootstrap Aggregating (Breiman 1996)). Bagging differs sharply from this heterogeneous partitions scenario. With bagging, multiple models are learned on random subsets of *overlapping* data *sampled from the general population*; the subsets are homogeneous, and the outputs of the ensemble are combined by simple voting.

In contrast, here, the MoE method must cope with non-overlapping, disjoint subsets, none of which are homogeneous in nature, and none of which share the same distribution as the general population. A future study motivated by this finding would be to compare MoE to a traditional ensemble method, such as Bagging, where the ensemble size is unconstrained. The idea would be to better understand the accuracy advantages accrued from MoE's local analysis versus the general accuracy improvements derived from ensemble methods.

Homogeneous Partitions Scenario

The right section of Table 2 presents the experimental results from the homogeneous partitions scenario. In this scenario, models are trained on equal-sized, homogeneous subsets of data, randomly sampled from the general population. There is no overlap in the data, however: each data instance in the general population is assigned to only one specific partition.

Performance of MoE Method These results paint a clear picture. First, as with the heterogeneous partitions scenario, the local MoE method statistically significantly outperformed the two global approaches (Weighted and Unweighted Average); this holds true across all datasets. However, while statistically significant, the advantage of the MoE approach over the global approaches was much less pronounced here than for heterogeneous partitions. The reasoning is straightforward; the density models are much more general and yield only minimal variation in the resulting weights (mixing coefficients) regardless of the input.

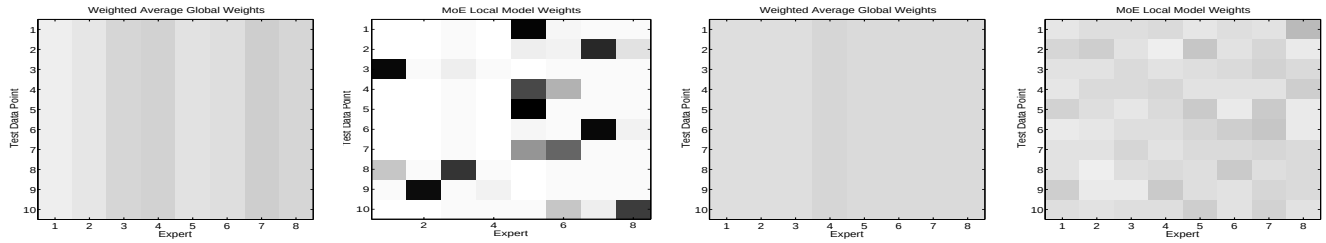
This follows, since the density models trained on a random sample of the general population will not have widely varying response for points in that population. Moreover, this finding is supported by Fig. 3d, which does show weights varying locally as expected, but only minimally so. Yet, the weights were still meaningful enough to outperform the Weighted and Unweighted Average approaches. In comparing Figs. 3c and 3d, it is clear that the any performance difference will be small since the weights approach those used for the Unweighted Average (i.e., uniform weighting).

Overall, the MoE method performed worse than the two Sages in the homogeneous partitions scenario, although performance was generally close.

Performance of Global Approaches There was no statistically significant difference in the Unweighted versus Weighted Averages approaches in the homogeneous partitions scenario. This is reasonable, since the weights are derived from classifier confidence, which is expected to be more or less equal across experts that were trained on partitions of similar data (the scenario considered here). This finding is illustrated in Fig. 3c; here, there is minimal if any variation in the global weights assigned to the eight experts. As a result, with uniform weights, output simply degenerates to that of the Unweighted Average.

Summary

These results provide evidence for the efficacy of the Mixture of Experts method under all scenarios involving disjoint data partitions. If the partitions are heterogeneous, the MoE approach using specialized local experts can exploit this to meet or even exceed the would-be performance of a single model having the advantage of being trained on all of the data. If the partitions are homogeneous, the local MoE method still performs better than other non-local methods. Regardless of the degree of partition homogeneity, if data are only available in disjoint partitions, the proposed local MoE method generally improves on naive global combiners, and in all cases, does no harm.



(a) Weighted Average Weights, Heterogeneous Partitions (b) Local MoE Weights, Heterogeneous Partitions (c) Weighted Average Weights, Homogeneous Partitions (d) Local MoE Weights, Homogeneous Partitions

Figure 3: Comparison of weight matrices for 8 experts over a random 10 instances in the *Adult* dataset, Weighted Average and Local MoE methods. Heterogeneous scenarios are shown at left; homogeneous at right. Lighter intensity indicates low weight for this model at this test point; darker intensity indicates higher weight.

5. Conclusions

In this paper, we adopted a local Mixture of Experts (MoE) method for application to disjoint inhomogeneous data partitions, that is, subsets of data whose distribution differs among the subsets and also from the general population. The central challenge under these circumstances is that models trained on one partition may not be applicable to (and may perform poorly on) test instances from other regions in feature space.

The statistical evaluation yielded two principal results. First, overall, the local MoE method performed on par with single-model and multiple-model "Sages" learned on all of the data; the class-conditional single Gaussian density model approach for estimating applicability is effective.

The second key result is that for both heterogeneous partitions and homogeneous partitions scenarios, for all datasets, the local MoE method outperformed the global Unweighted and Weighted Average methods. In such conditions, the experimental results never showed a penalty for combining experts according to their estimated local accuracy. If data are only available in disjoint partitions, then regardless of the degree of partition homogeneity, the proposed local MoE method generally helps versus naive global combiners, and in any case, never does harm.

Future Work

The first area for future work is an enhancement to this technique resulting in more elaborate density models. Instead of a single class-conditional Gaussian model, a Gaussian Mixture Model (GMM), separate from the MoE mixture model that combines the experts, could be used to estimate the distribution of the training data with additional fidelity.

Second, this study adopted an *equal-width* (distance-based) partitioning scheme described in (Han 2005). Han also presents an alternative frequency-based approach known as *equal-depth* partitioning. Evaluating the local MoE method's performance on inhomogeneous partitions generated by this alternative partitioning scheme will be useful.

Finally, we found that on two datasets in this study, the local MoE method performed statistically significantly better than the multiple-model Sage trained using Bagging. A more in-depth comparison of the MoE "disjoint subsets" ensemble technique versus Bagging's "overlapping subsets" ensemble approach will be a useful future study.

References

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

Cevikalp, H., and Polikar, R. 2008. Local classifier weighting by quadratic programming. *IEEE Trans. on Neural Networks* 19(10):1832–1838.

Chang, C.-C., and Lin, C.-J. 2001. *LIBSVM: a library for support vector machines.*

Han, J. 2005. *Data Mining: Concepts and Techniques.* San Francisco, CA: Morgan Kaufmann.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural Comput.* 3(1):79–87.

Jacobs, R. A. 1995. Methods for combining experts' probability assessments. *Neural Comput.* 7(5):867–888.

Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6:181–214.

Kusnezov, D. F. 2004. Advanced Simulation & Computing: The next ten years. Technical Report NA-ASC-100R-04, Sandia National Laboratories, Albuquerque, NM.

Moody, J., and Darken, C. J. 1989. Fast learning in networks of locally-tuned processing units. *Neural Computation* 1(2):281–294.

Procopio, M. J.; Kegelmeyer, W. P.; Grudic, G.; and Mulligan, J. 2009. Terrain segmentation with on-line mixtures of experts for autonomous robot navigation. In *Multiple Classifier Systems*, 385–397.

Ramamurti, V., and Ghosh, J. 1999. Structurally adaptive modular networks for non-stationary environments. *IEEE Trans. on Neural Networks* 10:152–160.

Sato, M.-A., and Ishii, S. 2000. On-line EM algorithm for the normalized Gaussian network. *Neural Computation* 12(2):407–432.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Woods, K.; Kegelmeyer, W.; and Bowyer, K. 1997. Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. Pattern Analysis and Machine Intelligence* 19(4):405–410.