A MIXTURE OF EXPERTS BASED DISCRETIZATION APPROACH FOR CHARACTERIZING SUBSURFACE CONTAMINANT SOURCE ZONES

Bilal Ahmed[1], *Itza Mendoza-Sanchez*[2], *Roni Khardon*[1], *Linda Abriola*[2], *Eric L. Miller*[1,3]

[1] Department of Computer Science, Tufts University
[2] Department of Civil and Environmental Engineering, Tufts University
[3] Department of Electrical and Computer Engineering, Tufts University
{bilal.ahmed,itza.mendoza-sanchez,roni.khardon,linda.abriola,eric.miller}@tufts.edu

ABSTRACT

Accidental releases and improper disposal of hazardous chemicals has led to widespread chemical contamination of subsurface soils and water-bearing formations. Effective remediation and restoration of such contaminated sites is dependent upon knowledge of the contaminant's mass and distribution within the aquifer. Recent research has shown that the estimation of certain metrics which summarize the distribution of the contaminant in the source-zone is sufficient for designing effective remediation strategies. In this work we explore the task of predicting such a metric based upon down-gradient concentration profiles. Motivated by the underlying physics of this problem we model this as a classification task where each class represents a particular sub-range of the metric. The solution to this problem is obtained by adapting the mixture of experts (MoE) scheme to learn a suitable quantization of the metric. Experimental evidence shows that this scheme outperforms baseline methods.

Index Terms— Subsurface Contamination, DNAPL Remediation, Source-Zone Characterization, Mixture of Experts, Classification

1. INTRODUCTION

The presence of hazardous chemicals such as dense non-aqueous phase liquids (DNAPLs) in the Earth's subsurface represents a significant potential threat for polluting groundwater resources and drinking water supplies. DNAPLs, including chlorinated organic solvents, are of particular concern, based upon their ubiquitous use in commercial and industrial products, large environmental releases, human toxicity, and tendency to persist for decades in the subsurface environment. Such spills tend to be distributed as mixtures of pools (localized areas of high contaminant saturation) and ganglia (more diffuse regions of much lower saturation). Work in [1] suggests that knowledge of *metrics* summarizing the pool and ganglia distribution may be sufficient for predicting the performance of a remediation strategy. Except for our preliminary work [2] there has not been much research on methods for deterimining such metrics from data available in the field.

Here we apply machine learning methods predicting the volume of DNAPL residing in pools at a given time relative to the initial spill volume, ρ_p. This metric is to be determined from observations of contaminant concentration observed in a transect oriented orthogonal to the nominal direction of groundwater flow located downgradient from the source zone, as well as a set of training data comprised of field-scale numerical simulations of realistic DNAPL source zone distributions and their evolving plumes under a variety of release and site heterogeneity conditions [1].

Given this information, an obvious approach to solve the problem would be the construction of a regression function to predict the metric value based upon the observed concentration profile for a particular source zone. It is an unfortunate fact, however, that the flow and transport physics mapping contaminant saturation into downstream concentration information is highly smoothed resulting in a substantial loss of information. If we think about source zone characterization from an inverse problems perspective, the problem would be very ill-posed. In terms of regression then, there is a concern that the error bounds associated with the point estimates of the source zone metric will be quite large. Therefore, we seek an alternative to regression and formulate this task as a classification problem in which the interval over which the metric is defined is quantized and the concentration data are used to determine the metric "bin" for a given source zone.

The formulation of the classification task requires that we quantize the metric into a finite number of non-overlapping levels. These levels define the classes for the subsequent classification of the observed concentration data. The learning task then entails that we have a procedure which jointly optimizes over the quantization of the metric and the degree of discrimination amongst the resulting classes. In order to solve this we adapt the mixture of experts (MoE) [3] scheme which

This work was supported by the Strategic Environmental Research and Development Program Project ER-1612 and by NSF grant IIS-0803409.

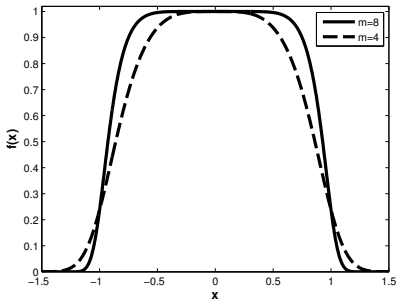

Fig. 1: The function $h(x, \mu, \beta) = 1 - \tanh((\frac{x-\mu}{\beta})^m)$ for $m = 4, 8$, $\mu = 0$ and $\beta = 1$.

divides the input space into multiple "regions" and the prediction for each region is done by a different "expert" [3]. In our case this corresponds to partitioning the feature space of the concentration data and then learning a probability distribution over the metric values in that partition. Our work modifies the intention of MoE because our primary goal is to identify the regions of expertise.

The rest of the paper is organized as follows: Section 2 provides the details of the MoE discretization scheme. In Section 3 we discuss our feature construction method and provide the results of our proposed scheme on a dataset comprised of field-scale numerical simulations of realistic DNAPL source zone distributions [1]. Finally, we conclude with discussion about the future directions of this work in Section 4.

2. MIXTURE OF EXPERTS FOR METRIC DISCRETIZATION

In a supervised learning task with an available training dataset $\mathcal{T} \equiv \{(x_i, y_i)\}$ where $x_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for regression or $y_i \in \{1, 2 \cdots, C\}$ for classification, the MoE model is:

$$P(y|x) = \sum_k P(z = k|x) P(y|x, z = k) \, . \tag{1}$$

The MoE model can be best understood as a conditional mixture model, i.e., each component of the mixture is a conditional distribution conditioned on the input x [3]. In (1) the probability distribution $P(z = k|x)$ is called the gating unit [3] and produces the mixing coefficients whereas $P(y|x, z = k)$ is the kth conditional distribution also known as the "expert".

In order to apply the MoE model to our problem, we take $x_i \in \mathbb{R}^n$ be the ith, length n feature input vector in our training set and $y_i \in \mathbb{R}$ the associated source zone metric (in our case ρ_p). The goal here is to use our training data to determine "bins" for the metric along with a classifier such that test data are placed in the correct bins. Thus, in our case the expert models the distribution of metric values in a particular "bin" and is used to determine the discretization of the metric. Assuming we are using K bins, we define $t_i \in \mathbb{R}^K$ as the class-label vector of each concentration observation x_i where t_{ik} is one for the bin into which this datum belongs and zero for all other $K - 1$ entries. Given a training data set of N instances with the associated metric values, the task is to estimate the unknown label vector for each instance. The collection of instances for which the estimated label vectors are the same then determines the cluster of feature vectors associated with the bin. Similarly, the corresponding metric values for that group define the extent of the bin in metric space. In previous work we modeled bin distributions using Gaussians, which can lead to large overlap [2]. So, to obtain bins in metric space that minimally overlap, we model the distribution of metric values in the k-th bin using the following function:

$$P(y|x, z = k) = \\ f(y, \mu_k, \beta_k) = \frac{1}{\Omega_k} \left(1 - \tanh((\frac{y - \mu_k}{\beta_k})^m)\right) \tag{2}$$

where Ω_k is the normalization constant ensuring that $f(y, \mu_k, \beta_k)$ is a valid PDF, m is an even number greater than two, μ_k defines the center of the bin and β_k determines the width of the bin. This is a box-like function as shown in Figure 1. The overall data likelihood of the model can then be described as:

$$L = \prod_{i=1}^{N} \prod_{j=1}^{K} [g(x_i, w_k) f(y_i, \mu_k, \beta_k)]^{t_{ik}} \tag{3}$$

where we use a *soft-max* function for modeling the mixing coefficients

$$P(z = k|x) = g(x, w_k) = \frac{e^{w_k^T x}}{\sum_{l=1}^{K} e^{w_l^T x}} \, . \tag{4}$$

The model is thus determined by the parameters $\eta_j = \{w_j, \mu_j, \beta_j\}$, $\forall j = 1 \ldots K$. Determining these parameters along with the label vectors constitutes the training phase.

2.1. Learning the Parameters

To learn the model parameters because the class label vectors t_i are hidden, we make use of the Expectation-Maximization (EM) algorithm [3, 4]. The EM algorithm is an iterative optimization approach where each iteration is comprised of two steps. In the "E" step we compute the posterior probabilities of the hidden class labels using the current estimate of the model parameters as:

$$\gamma_{ik} = \frac{g(x_i, w_k) f(y_i, \mu_k, \beta_k)}{\sum_{l=1}^{K} g(x_i, w_l) f(y_i, \mu_l, \beta_l)} \, . \tag{5}$$

The expectation of the model's log-likelihood function with respect to the posterior distribution over the latent variables can be written as:

$$Q(\eta, \eta^{(p)}) = \sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ij} \left[\ln(g(x_i, w_j)) - \ln(\Omega_j) \right. \\ \left. + \ln(h(y_i, \mu_j, \beta_j))\right] \tag{6}$$

where $h(y_i, \mu_j, \beta_j) = 1 - \tanh((\frac{y_i - \mu_j}{\beta_j})^m)$ and $\eta^{(p)}$ are the parameter values at iteration p. In the "M" step we estimate the new parameter values $\eta^{(p+1)}$ by maximizing this function with respect to η.

2.1.1. Updating the Gating Unit Parameters

The Q function (6) is influenced by the parameters of the gating unit only through the terms $\gamma_{ij} \ln(g(x_i, w_j))$. Thus, in order to maximize the Q function with respect to the gating unit parameters we solve the following maximization problem:

$$w_k^{(p+1)} = \arg\max_{w_k} \sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ij} \ln(g(x_i, w_j))$$

This is a maximum likelihood formulation for a multinomial logisitic regression problem with the otuputs defined by γ_{ij} [4]. This can be solved by using the iterative reweighted least squares (IRLS) algorithm [4].

2.1.2. Updating the Bin Parameters

In order to estimate the new values of the binning function parameters $\nu_k = \{\mu_k, \beta_k\}$ we can formulate the following maximization problem by noting the dependancy of the Q function (6) on these parameters:

$$\nu_k^{(p+1)} = \arg\max_{\mu_k, \beta_k} \sum_{i=1}^{N} \sum_{j=1}^{K} \gamma_{ij} \left[\ln(h(y_i, \mu_j, \beta_j)) - \ln(\Omega_j)\right]$$

This is a non-linear unconstrained maximization problem which we solve using the BFGS algorithm [5] which requires the computation of the gradient of the above function with respect to the individual parameters. The derivatives of the objective function with respect to the parameters are:

$$\frac{\partial Q}{\partial \mu_k} = \frac{m}{\beta_k} \sum_{i=1}^{N} \gamma_{ik} \phi_{ik}^{m-1} (1 + \tanh(\phi_{ik}^m)) \quad (7)$$

$$\frac{\partial Q}{\partial \beta_k} = \frac{1}{\beta_k} \sum_{i=1}^{N} \gamma_{ik} \left[m \phi_{ik}^m (1 + \tanh(\phi_{ik}^m)) - 1\right] \quad (8)$$

where $\phi_{ik} = (y_i - \mu_k)/\beta_k$.

2.1.3. Testing

Once we have the final estimates of the model parameters, we can test our model on a previously *held-out* subset of data. For a test instance (x_i, y_i) we consider the correct bin to be $\text{argmax}_k\ f(y_i, \mu_k, \beta k)$. Thus, although the MoE does not guarantee disjoint bins we force the bins to be disjoint by comparing the corresponding PDFs. The predicted bin is $\text{argmax}_k\ g(x_i, w_k)$, and the instance is considered to be correctly classified if both choose the same label.

3. EXPERIMENTAL ANALYSIS

3.1. Morphological Feature Construction

The features we use for processing are motivated by our intuition concerning how the morphology of the concentration data images is related to that of the DNAPL saturation in the source zone. In the down-gradient concentration profiles the regions of high concentration are typically associated with pool-like saturation distributions, while in the ganglia dominated regions of the source-zone the corresponding regions in the concentration images are more diffuse. Thus we seek features that capture the size and number of "blobs" in the concentration data believing that they are related to the characteristic metric. We specify the "blobs" at some level τ to be those pixels in the image whose concentrations exceed τ; i.e., $b(x, y; \tau) = 1$ if $c(x, y) > \tau$ and is zero otherwise; where $c(x, y)$ is the observed concentration image at pixel location (x, y). From $b(x, y; \tau)$ we compute two quantities: the percentage of the area in $c(x, y)$ for which $b(x, y; \tau) = 1$ and the number of connected components at that level. The percentage of area calculation is $\pi(\tau) = \sum_{x,y} b(x, y; \tau) / \sum_{x,y} b(x, y; 0)$ where the denominator is the number of pixels in the concentration image that are nonzero. We denote by $\nu(\tau)$ the number of connected components at a threshold value of τ. The morphological feature vector we create is comprised of $\pi(\tau)$ and $\nu(\tau)$ for $\tau = 0, 1, 2, \cdots, \tau_{max}$ where $\tau_{max}$is the largest value of concentration in the training data set.

3.2. Results

We demonstrate the effectiveness of our proposed discretization scheme on a hydrological dataset gathered from Sequential Gaussian Simulation of the subsurface comprised of 593 instances. There are a number of parameters which control the nature and behavior of the contaminant in the source-zone including volume of the contaminant spilled, the release rate of the contaminant during the spill and the physical area over which the contaminant was spilled. In our simulations we have varied the spill-rate between 4 and 400 days and have also used different release configurations (physical locations of the injection points).This makes the task of predicting the source-zone metrics more challenging and also provides more data diversity to our learning algorithm. To test the performance of our model we compare its results to two other discretization strategies. In uniform binning [6] the entire range of the continuous valued attribute is divided into k bins of equal length. Similarly, in equal-frequency binning [6] the entire range of the continuous valued attribute is divided into k bins such that the frequency of samples in each bin is uniform.

The comparison of the performance of the three approaches is shown in Table 1. We used a ten-fold cross-validation scheme for measuring the performance of the three approaches. To

Table (a)	Bin-1	Bin-2	Bin-3	Bin-4
Range	$[0, 0.08]$	$[0.08, 0.43]$	$[0.43, 0.81]$	$[0.81, 1]$
No. Observations	158	142	102	191
Accuracy	0.82 ± 0.10	0.69 ± 0.10	0.73 ± 0.12	0.92 ± 0.05

Table (b)	Bin-1	Bin-2	Bin-3	Bin-4
Range	$[0, 0.25]$	$[0.25, 0.5]$	$[0.5, 0.75]$	$[0.75, 1]$
No. Observations	229	87	64	213
Accuracy	0.84 ± 0.08	0.49 ± 0.18	0.56 ± 0.25	0.91 ± 0.06

Table (c)	Bin-1	Bin-2	Bin-3	Bin-4
Range	$[0, 0.06]$	$[0.06, 0.41]$	$[0.41, 0.9]$	$[0.9, 1]$
No. Observations	148	148	148	149
Accuracy	0.76 ± 0.11	0.66 ± 0.13	0.72 ± 0.11	0.79 ± 0.1

Table 1: Results of 10-fold cross-validation for the three approaches. 90% of the observations were used for training while the remaining 10% were retained as test instances and used to test the accuracy of the classifier. The discovered bin ranges and their classification accuracies: (a) using the MoE model, (b) using uniform binning strategy, and (c) using equal frequency binning strategy.

calculate the predictions and measure the performance of the two baseline strategies we used multiclass logisitic regression. The bin-ranges shown in Table 1(a) represent an average across the bin-boundaries obtained using the ten-fold cross-validation results for the MoE model. As the results show the two baselines produce higher accuracies in different regions. The uniform binning strategy produces higher accuracies for the first and the last bin which correspond to low and high metric values and performs poorly in the middle two bins. On the other hand the equal-frequency binning methodology outperforms the unifrom binning methodology in the middle two bins. The MoE approach performs better than either of the two baseline techniques in the two middle bins and as well as the two baselines in the lower and higher metric range.

4. CONCLUSIONS AND FUTURE WORK

In this work we have taken a different approach to modeling the prediction of metric values that characterize source-zone architecture from down-gradient plume responses. Owing to the dearth of information in the response signals we have modeled this as a classification problem and instead of providing exact regression estimates we provide a range estimate for the metric value with high accuracy. We have used the MoE model to estimate a suitable discretization of the metric and in this work we have specifically shown that it can be effectively used for predicting the percentage-mass of DNAPL in pools from the down-gradient concentration profiles. Comparing to our preliminary work [2] where we used Gaussians to model the bin distribution, we get similar accuracies and bin ranges but the overlap between bins is considerably smaller. In the future we would like to compare the effectiveness of this approach against more sophisticated methods that explicitly force bins to be disjoint and also measure its performance against regression based on the traditional MoE model. Another avenue worthy of more research is to test the robustness of this approach with regards to its peformance on various other source-zone metrics.

5. REFERENCES

[1] John A. Christ, Andrew C. Ramsburg, Kurt D. Pennell, and Linda M. Abriola, "Predicting DNAPL mass discharge from pool-dominated source zones," *Journal of Contaminant Hydrology*, vol. 114, pp. 18–34, 2010.

[2] Bilal Ahmed, Roni Khardon, Itza Mendoza-Sanchez, Linda M. Abriola, and Eric L. Miller, "A discriminative-generative approach to the characterization of subsurface contaminant source zones," *To appear in the Proceedings of the International Geoscience and Remote Sensing Symposium*, 2012.

[3] Michael I. Jordan and Robert A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, March 1994.

[4] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 3 edition, 2006.

[5] Jorge Nocedal and Stephen J. Wright, *Numerical Optimization*, New York: Springer-Verlag, 2006.

[6] James Dougherty, Ron Kohavi, and Mehran Sahami, "Supervised and unsupervised discretization of continuous features," in *Proceedings of the Twelfth International Conference on Machine Learning*, 1995, pp. 194–202.