A DISCRIMINATIVE-GENERATIVE APPROACH TO THE CHARACTERIZATION OF SUBSURFACE CONTAMINANT SOURCE ZONES

Bilal Ahmed[1], *Itza Mendoza-Sanchez*[4], *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
[4] Escuela Superior de Ingenieria y Arquitectura, Mexico City, Mexico
{bilal.ahmed,itza.mendoza-sanchez,roni.khardon,linda.abriola,eric.miller}@tufts.edu

ABSTRACT

Large-scale contamination of ground water due to improper disposal of hazardous chemicals poses a global threat to drinking water supplies. Effective restoration and remediation of such sites relies upon a knowledge of the contaminant's distribution within the subsurface. Obtaining a detailed map of the existing distribution is usually not feasible; rather partial knowledge in terms of certain metrics that characterize the distribution has recently been shown to be sufficient for planning and monitoring remediation strategies. In this work we explore the prediction of a representative metric based upon down-gradient concentration profiles using a classification framework where each class represents a particular sub-range of the metric. Initial experiments show that our proposed model can be used effectively for predicting the metric.

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

1. INTRODUCTION

Accidental releases and improper disposal of hazardous chemicals has led to widespread chemical contamination of subsurface soils and water-bearing formations, threatening groundwater resources and drinking water supplies. Dense non-aqueous phase liquids (DNAPLs), such as chlorinated organic solvents, are of particular concern, based upon their ubiquitous use in commercial and industrial products, very large environmental releases, human toxicity, and tendency to persist for decades in the subsurface environment. A key component in the design of a remediation strategy for these sites is the characterization of the *source zone architecture*, that is the existing distribution of DNAPL in the subsurface. Recent work in [1] suggests that a detailed map of the subsurface may not be required to predict the performance of a remediation strategy. Rather, *metrics* characterizing the distribution of contaminant may suffice. Specifically, in [1], the authors demonstrated that the ratio of the volume in the source zone occupied by ganglia (DNAPL saturation below a specified threshold) to that of pools (saturation exceeding the threshold) could be used within a model of subsurface flow and transport to predict remediation performance. In that work however, the authors did not consider how this ganglia-to-pool ratio (GTP) could be determined from data available in the field.

Motivated by [1], in this work we explore the use of machine learning methods for determining metrics defining the source zone architecture. We are interested in predicting percentage-mass of DNAPL in pools (ρ_p) which defines the volume of DNAPL residing in pools at a given time relative to the initial spill volume. This metric is to be determined based upon observations of contaminant concentration observed in a transect oriented orthogonal to the nominal direction of groundwater flow 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, one approach to solve the problem would be the construction of a regression function using any of a number of statistical or machine learning methods [2]. Motivated by a concern that the nature of the flow and transport physics associated with this problem will result in unacceptably large confidence intervals in regression processing, here we formulate and solve a classification problem in which the intervals over which the metric is defined are quantized and the concentration data are used to determine the metric "bin" for a particular source zone. The hope here is that the widths of the bins will be smaller than typical confidence bounds for the regression approach. The answer to this question as well as determining how quantized metric estimates can be used in simplified flow and transport models

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

to predict remediation performance are issues currently under investigation. Here we are concerned only with the problem of constructing a classifier to solve the problem just described.

The problem as posed above has two main components: a quantization of the metric and subsequent classification of the concentration observations into the discovered metric "bins". These two problems are coupled as we need to find a quantization of the metric that provides high classification accuracy based on the discovered quantization. From a machine learning perspective we seek groups of concentration observations which are easily discernible from each other and at the same time possess (possibly) non-overlapping ranges of the metric. To address this problem we employ a *generative* model [2] that explicitly models generation of bins and generation of values in bins. The bins and their values are discovered using the iterative expectation maximization (EM) algorithm [2]. At each iteration the algorithm produces a quantization of the metric into k bins. Using these bins as class-labels a *discriminative* multi-class logistic regression model is trained to estimate the separability of the k classes. Based on the output of the classifier the bins are readjusted until the EM algorithm converges. This provides us with a mixture of experts (MoE) [3] type of algorithm which fuses the two frameworks of supervised and unsupervised learning for quantizing the regression output. The mixture of experts (MoE) scheme divides the input space into multiple "regions" and then learns a separate regression or classification function termed an "expert" in each region. The prediction for a new instance is estimated as a weighted combination of the output of all the experts [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 only to identify the regions of expertise which in turn define the range of each bin. Section-2 provides the details of our proposed metric quantization scheme.

One practical problem we face in this application is the relatively small amount of data available for learning. Indeed, the computational complexity of simulating the underlying physics allows for producing datasets on the order of a few hundred examples. Hence, we first subject the concentration profiles (each of which comprises of thousands of pixels) to a feature extraction stage. We use morphological signal processing methods to extract more relevant information from the raw concentration profiles [4]. This transformed data is then subjected to a standard dimension reduction method, principal component analysis (PCA) to overcome the *curse-of-dimensionality* [2]. To test the efficacy of our algorithm we compare its performance to uniform binning and equal frequency binning strategies [5]. In uniform binning the range of the continuous valued attribute is divided into k bins of equal length and in the case of equal frequency binning the range is divided into k bins such that all bins contain the same number of data points. Section-3 explains our feature construction methodology and provides the classification results. Finally we conclude with a discussion of future research directions in Section-4.

2. A DISCRIMINATIVE-GENERATIVE APPROACH TO METRIC QUANTIZATION

In the context of estimating source zone metrics, we take $r_i \in \mathbb{R}^n$ to be the ith, length n feature input vector in our training set and $m_i \in \mathbb{R}$ 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. Assuming we are using K bins, we define $t_i \in \mathbb{R}^K$ as the class-label vector of each concentration observation r_i where t_{ij} 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. Collecting those 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. We model the distribution of metric values in the k-th bin as a Gaussian random variable with mean μ_k and variance σ_k. Determination of these parameters, which basically define the distribution of the metric values, along with the label vectors constitutes the training phase of this approach to metric quantization.

2.1. Discriminative-Generative Model

Our model is based on the multi-class formulation of logistic regression which is coupled with a normal distribution for each bin. Using this model the likelihood function is given by:

$$L = \prod_{i=1}^{N} \prod_{k=1}^{K} \left[y_{ik} N(m_i | \mu_k, \sigma_k^2) \right]^{t_{ik}} \tag{1}$$

In (1) y_{ik} models the class conditional distribution $p(t_{ik} = 1|r_i)$ for the i-th instance of training data belonging to the k-th class and is formulated using the "soft-max" function:

$$p(t_{ik} = 1|r_i) = y_{ik} = \frac{e^{w_k^T r_i}}{\sum_{l=1}^{K} e^{w_l^T r_i}}. \tag{2}$$

The model described in (1) looks for groups of data in the feature space which are easily separable from each other and at the same time have tightly clustered metric values. As can be seen from (1) class conditional probabilities of the observations are directly coupled with the probabilities of the associated metric values. Thus (1) would be maximized for groups of observations which are easily identifiable within the logisitic regression (classification) framework and are also tightly grouped with respect to the metric ranges they span.

2.2. Learning the model paramters

In the model described above the concentration observations r_i and the associated metric values m_i are observed, but the label vectors t_i are not observed. The parameters for the model are $\{w_k, \mu_k, \sigma_k^2\}$. To learn these parameters, we make use of the Expectation-Maximization (EM) algorithm commonly employed for problems of this type [2]. The EM algorithm is an iterative optimization approach where each iteration is comprised of two steps: the computation of the posterior probabilities of the latent variables in the first step and a certain expected value which is then maximized with respect to the model parameters in the second step. In the first "E" step we compute the posterior probabilities of the hidden class labels using the current estimate of system parameters as:

$$\gamma_{ik} = \frac{y_{ik} N(m_i|\mu_k, \sigma_k^2)}{\sum_{l=1}^{K} y_{il} N(m_i|\mu_l, \sigma_l^2)} \tag{3}$$

For the M-step we replace the t_{ik} with γ_{ik} and maximize the expected value of the complete-data log-likelihood (where the expectation is taken under $P(t_{ik}|r_i, \theta^{old})$) with respect to the parameters. The EM updates for the mean and variance parameters are given in closed form as follows:

$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{ik} m_i}{\sum_{j=1}^{N} \gamma_{jk}} \quad , \quad \sigma_k^2 = \frac{\sum_{i=1}^{N} \gamma_{ik} (m_i - \mu_k)^2}{\sum_{j=1}^{N} \gamma_{jk}}$$

In order to update the k-th weight vector we need to optimize:

$$Q(w) = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \ln y_{ik} \tag{4}$$

Note that (4) is the same as a likelihood function for a simple multi-class logistic regression model with the labels replaced by γ_{ik}. We have used the iterative reweighted least squares (IRLS) algorithm [6] for solving this optimization problem. The IRLS algorithm is based on a Newton-Raphson update and requires the computation of the Hessian matrix [6]. Thus, in our formulation once the parameters have been initialized we can run the simple multi-class logistic regression model to find the estimates for the weights required in the M-step by simply treating all the γ_{ik} as true labels.

2.2.1. 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 (r_i, m_i) we consider the correct bin to be the one that has the maximum probability for m_i as

$$k = \arg\max_{k} \; N(m_i|\mu_k, \sigma_k^2)$$

The predicted bin is calculated using (2), and the instance is considered to be correctly classified if both choose the same bin. Thus, although conventional MoE does not guarantee disjoint bins we force the bins to be disjoint by comparing the corresponding PDFs.

3. EXPERIMENTAL ANALYSIS

We demonstrate the effectiveness of our proposed discretization scheme on a hydrological dataset gathered from DNAPL infiltration and entrapment simulations in permeability fields generated with the sequential Gaussian geostatistical method comprising 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 0.32 and 32 L/day and have also used different release configurations (physical locations of the injection points). This makes the task of predicting the source-zone metric more challenging and also provides more data diversity to our learning algorithms. Some of the data used here has been reported [1] for characterizing subsurface contaminant source-zones.

3.1. Feature Construction

The raw data used as the basis for classification are observations of contaminant concentration collected in a transect orthogonal to groundwater flow located some distance from the source zone. We have found [4] that morphology of "blobs" in this concentration image is in fact related to the upstream source zone architecture. More formally, the feature vector used in our experiments is constructed in two steps. The first step is a tresholding procedure, where 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

$$\pi(\tau) = \frac{\sum_{x,y} b(x, y; \tau)}{\sum_{x,y} b(x, y; 0)}$$

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 τ_{max} is the largest value of concentration in the training data set. In the second step, PCA is applied to these vectors to obtain a six dimensional linear transformation that is used to compute feature vectors that are input to the algorithm.

Table (a)	Bin-1	Bin-2	Bin-3	Bin-4
Range	$[0, 0.07]$	$[0.07, 0.43]$	$[0.43, 0.82]$	$[0.82, 1]$
No. Observations	156	144	106	187
Accuracy	0.82 ± 0.09	0.68 ± 0.10	0.78 ± 0.12	0.92 ± 0.07

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 100-fold random cross-validation for the two approaches on the first dataset. 90% of the observations were used for determining bin ranges and training a multi-class logistic regression classifier 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 using (a) the discriminative-generative model, (b) the uniform binning strategy, and (c) equal frequency binning strategy.

3.2. Results

In order to test the performance of our proposed discretization scheme, we employed a 100-fold random cross-validation strategy. At each iteration 90% of the data were used for training and 10% for testing with the partition determined randomly. The results of the algorithm along with those for the uniform and equal frequency binning strategies are shown in Table-1. The proposed algorithm clearly outperforms the uniform binning strategy in the two middle bins and produces comparable results in the other two bins. Whereas, compared to equal frequency binning our algorithm outperforms it in the first and the last bin. It therefore produces more consistent and higher accuracies throughout the entire range of the metric as compared to the two baseline techniques.

4. CONCLUSIONS AND FUTURE WORK

We have adapted the MoE scheme for a new task, discretizing the output variable for regression tasks where the input data may not be suitable for calculating point-estimates. The initial results of the proposed algorithm on hydrological data show that it can be affectively used for predicting the percentage-mass of DNAPL in pools from down-gradient concentration profiles. In more recent work [7] we have extended this scheme to use box-like distributions that are more suitable for non-overlapping bins than the normal distributions used in this paper. As future work we would like to explore the performance of our method in the estimation of a wider range of source zone using more comprehensive sets of data. We would also like to incorporate kernel based methods within this discriminative-generative framework for label discretization.

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

[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] Hao Zhang, Itza Mendoza-Sanchez, Linda M. Abriola, and Eric L. Miller, "Manifold Regression For Subsurface Contaminant Characterization," *To appear in the Proceedings of the International Geoscience and Remote Sensing Symposium*, 2012.

[5] 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.

[6] Thomas P Minka, "Algorithms for maximum-likelihood logistic regression," *Statistics Tech Report, Carnegie Mellon University*, pp. 1–15, 2001.

[7] Bilal Ahmed, Itza Mendoza-Sanchez, Roni Khardon, Linda M. Abriola, and Eric L. Miller, "A mixture of experts based discretization approach for characterizing subsurface contaminant source zones," *To appear in the Proceedings of the IEEE Statistical Signal Processing Workshop*, 2012.