MIXTURE OF HMM EXPERTS WITH APPLICATIONS TO LANDMINE DETECTION

Seniha Esen Yuksel [a], *Paul D. Gader* [b]

[a] Department of Materials Science and Engineering, University of Florida
[b] Department of Computer & Information Science & Engineering, University of Florida

ABSTRACT

This paper introduces a novel mixture of experts model, the Mixture of Hidden Markov Model Experts (MHMME). This model is designed to perform context-based classification of samples that are variable length sequences. The contexts are determined by the gates and the classifiers are determined by the experts. The gates and the experts are learned simultaneously using a single probabilistic model. Experimental results on landmine dataset show that MHMME significantly outperforms the HMM-based and ME-based models.

Index Terms— Mixture of experts, hidden Markov models, ME, HMM, landmine detection, metal detector, WEMI.

1. INTRODUCTION

Finding contexts from data can significantly increase the classification rates when classes contain multiple interlaced subclasses whose characteristic are dependent on the context. However, finding context and classifier models simultaneously is very difficult if the observed data consists of sequences. One such example is in landmine detection. Landmines appear in many sizes and shapes and are roughly categorized into four groups according to their metallic content and intended targets as high metal anti-tank (HMAT), high metal anti-personnel (HMAP), low metal anti-tank (LMAT), and low metal anti-personnel (LMAP). However, these groups mostly overlap, and the signals collected from these mines can be significantly affected by changes in temperature, humidity, and soil conditions. Therefore, the contexts are generally hard to define, they are often interlaced, and do not have sharp boundaries. In such cases, we define a context as a group of similar signatures.

In this study, a novel mixture of hidden Markov model experts (MHMME) is developed that can both decompose sequential data into multiple contexts and learn expert classifiers for each context. In this model, a gate of HMMs defines the contexts and cooperates with a set of HMM experts that provide multi-class classification. The MHMME model is inspired from the mixture of experts (ME) model [1], and extends it to sequential (and time-series) data for classification. Therefore, MHMME carries the advantages of the ME model and also brings advantages that set it apart from the other models, summarized as below:

- MHMME model provides a divide and conquer approach, is probabilistic, and has soft boundaries – all of which support context learning. Unlike the traditional mixture models where the mixture coefficient is a scalar, in MHMME the mixture coefficient (i.e. the gate) depends on the input and helps define the contexts that are unknown to the data modeler.
- The learning of the contexts and the classifiers is accomplished simultaneously, in one model. During training, there is no hard clustering of data, which means that the sequences can freely move between contexts and classifiers.
- MHMME considers the temporal connections in time-series data, and is suitable for high-dimensional sequential data of varying lengths due to the use of the hidden Markov models (HMMs). In addition, HMMs at the gate and the experts can be of different topologies (number of states, observation symbols etc.).

In the ME literature, a number of models [2–6] were described that extend the ME architecture to time-series data. These models, however, are only applicable to regression, and they use a multi-step-ahead prediction in which the last values of the time-series data are used as features in a neural network. Such models cannot handle data of varying length and the use of multilayer network-type approaches prevent them from completely describing the temporal properties of a time-series dataset. In contrast to these models, our study is on classification and is suitable for varying length sequences.

2. MIXTURE OF HMM EXPERTS

For all the hidden Markov models, we define:

- W = number of states; M = number of symbols in the codebook; T = length of observation sequence.
- I = number of experts; K = number of classes.
- $V = \{v_1, ..., v_M\}$ the discrete set of observation symbols. $O = O_1 O_2 ... O_T$ denotes an observation sequence, where O_t

This research was partially supported by the Army Research Office grant W911NF0510067 and the NSF Grant No. 0730484.

is the observation at time t; $Q = q_1 q_2 ... q_T$ is a fixed state sequence, where q_t is the state at time t; $S = \{S_1, S_2, ..., S_W\}$ are the states.

- λ_{ik} = HMM model for the k^{th} class at the i^{th} expert such that $\lambda_{ik} = \{A^{(ik)}, B^{(ik)}\}$. Also, $\psi_i = i^{th}$ is the HMM model at the gate.
- The initial state distribution $\pi = \{\pi_r\}_{r=1}^{W}$, where $\pi_r = P(q_1 = S_r)$ is the probability of being in state r at time $t = 1$.
- The state transition probability $A = \{\{a_{rj}\}_{r=1}^{W}\}_{j=1}^{W}$, where $a_{rj} = P(q_{t+1} = S_j | q_t = S_r)$.
- The emission matrix $B = \{\{b_j(m)\}_{j=1}^{W}\}_{m=1}^{M}$, where $b_j(m) = P(v_m \texttt{ at } t | q_t = j)$.

The MHMME architecture is illustrated in Fig. 1 where the gate has I HMM models. Each branch of the gate is connected to an expert, and an expert has K HMMs, one for each class. The gate partitions the set of all time-series data that can serve as inputs to the HMMs, and defines the contexts where the individual expert opinions are trustworthy. Experts discriminate data in these partitions based on class labels. We denote the HMM models at the gate with $\Psi = \{\psi\}_{i=1}^{I}$, the HMM models at the experts with $\Lambda_i = \{\lambda_{ik}\}_{k=1}^{K}$, and finally, we denote the set of all the gate and expert parameters as $\Theta = \{\Psi, \Lambda\}$. Let the data be denoted by $D = \{\mathbf{O}, Y\}$ where $\mathbf{O} = \{O^{(n)}\}_{n=1}^{N}$ represents the input sequences, and $Y = \{\mathbf{y}^{(n)}\}_{n=1}^{N}$ represents the class coded true outputs of training data such that $\mathbf{y}^{(n)} = [y_1^{(n)}, ..y_k^{(n)}, ..y_K^{(n)}]$, and $y_k^{(n)} = 1$ if $x^{(n)}$ belongs to class k, and 0 otherwise. The gate and experts make a decision following the complete data distribution

$$P(D, Z; \Theta) = \prod_n \prod_i \left(g_i^{(n)} P_i(y^{(n)}) \right)^{z_i^{(n)}}$$

where $z_i^{(n)}$ is a latent variable, $g_i^{(n)} = P(i | O^{(n)}, \Psi_i)$ is the probability of selecting the i^{th} expert given the sequence $O^{(n)}$. $P_i(\mathbf{y}^{(n)})$ is the probability that the i^{th} expert has generated $\mathbf{y}^{(n)}$ given $O^{(n)}$. The gate's estimate is obtained by:

$$g_i^{(n)} = \frac{\exp f(O^{(n)} | \psi_i)}{\sum_{m=1}^{I} \exp f(O^{(n)} | \psi_m)}$$

where $f(O^{(n)} | \psi_i)$ is the Viterbi log-likelihood of observation $O^{(n)}$ for an HMM model ψ_i. Similar to the gate, the HMMs at the experts compute the Viterbi log-likelihood

$$f(O^{(n)} | \lambda_{ik}) = \log P_{HMM}(O^{(n)}, Q, \lambda_{ik})$$

where

$$P_{HMM}(O, Q, \lambda_{ik}) = \pi_{q_0}^{(ik)} \prod_{t=1}^{T-1} a_{q_t q_{t+1}}^{(ik)} \prod_{t=1}^{T} b_{q_t}^{(ik)}(o_t)$$

is the Viterbi likelihood. The output of expert i for class k is $\hat{y}_{ik}^{(n)}$, computed as

$$\hat{y}_{ik}^{(n)} = \exp f(O^{(n)} | \lambda_{ik}) / \sum_{r=1}^{K} \exp f(O^{(n)} | \lambda_{ir}).$$

Therefore, in the E step, we find the expectations $h_i^{(n)}$ of the hidden variables and in the M step, we find the HMM parameters that maximize the objective functions Q_e and Q_g:

$$\psi_i^{(p+1)} = \underset{\psi_i}{\text{argmax}}\, Q_g = \underset{\psi_i}{\text{argmax}} \sum_{n=1}^{N} \sum_{i=1}^{I} h_i^{(n)} \log g_i^{(n)} \quad (1)$$

$$\lambda_{ik}^{(p+1)} = \underset{\lambda_{ik}}{\text{argmax}}\, Q_e = \underset{\lambda_{ik}}{\text{argmax}} \sum_{n=1}^{N} \sum_{i=1}^{I} h_i^{(n)} \log P_i(\mathbf{y}^{(n)}) \quad (2)$$

To ensure that the estimated parameters satisfy the constraints $a_{rj} \geq 0$, $\sum_{j=1}^{W} a_{rj} = 1$, $b_{mj} \geq 0$, and $\sum_{m=1}^{M} b_{mj} = 1$, we map these parameters using log, and map them back with softmax functions as in [7, 8]. Let p denote the iteration number. The HMM parameters that maximize the objective functions are found by gradient ascent updates as:

$$\tilde{a}_{rj}^{(ik)}(p+1) = \tilde{a}_{rj}^{(ik)}(p) + \epsilon \frac{\partial Q_e(\Lambda(p))}{\partial \tilde{a}_{rj}^{(ik)}(p)},$$

and

$$\tilde{b}_{mj}^{(ik)}(p+1) = \tilde{b}_{mj}^{(ik)}(p) + \epsilon \frac{\partial Q_e(\Lambda(p))}{\partial \tilde{b}_{mj}^{(ik)}(p)}.$$

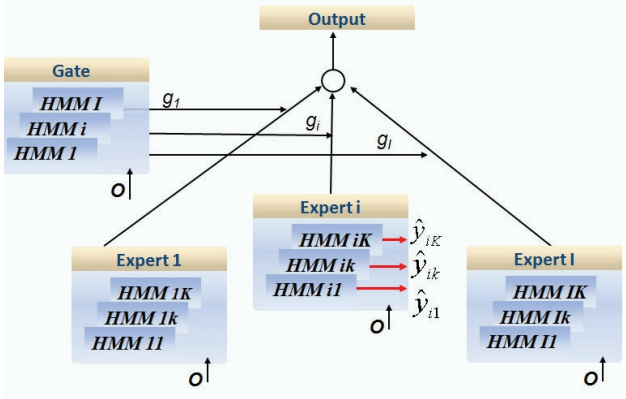


Fig. 1. MHMME architecture with I experts for K classes. A gate partitions the set of all sequential or time-series data that can serve as inputs to the HMMs. Experts learn to discriminate the classes in these partitions.

3. EXPERIMENTS ON LANDMINE DATA

For landmine detection we consider a dataset consisting of mine and non-mine object data collected using a robotic system with wide-band electromagnetic induction (WEMI) sensors [9, 10]. The data were collected from a controlled environment and was described in [11, 12]. The WEMI sensors

collect complex responses in 21 frequencies between 330Hz and 90, 030Hz which can be modelled as $S(w) = A[I(w) + iQ(w)]$ where w is the frequency, A is the magnitude, $I(w)$ is the real and $Q(w)$ is the imaginary response of the complex system. The term $I(w) + iQ(w)$ describes the shape of the response, and it can be represented by an Argand diagram which is the plot of $I(w)$ with respect to $Q(w)$. The shape of an Argand diagram is indicative of the type and distribution of metal in a target, and mines of the same type show similar Argand curves that are scaled replicas of each other depending on the depth. However, the features from the mines are generally interlaced and it is difficult to appoint a model as an expert to identify a particular subclass of mines [13, 14]. Therefore the MHMME model can find multiple models from the data that represent each of these contexts, and do a better classification than those ignoring the context.

The data was normalized between $[0, 1]$ to eliminate the variation in magnitude due to depth, and the Argand sequences were discretized to the 50 cluster centers found by FCM clustering. The MHMME architecture was set to have 8 experts, which corresponds to 8 HMMs at the gate, and 2 HMMs at each expert. All the HMMs were set to have 3 states. To initialize the gate, 4 HMMs were learned from class 1 (mines) and 4 HMMs were learned from class 2 (non-mines) using Cl-HMM [15]. Then, all the training sequences were tested with all the HMMs and those that received high likelihoods were used to initialize an HMM for each expert using Baum-Welch training. The sequences with the highest probabilities at each gate represent the contexts defined by the gate. These sequences are displayed in Fig. 2 where the type of the mine or nonmine object is given in the y-axis. For example, the first context is defined by the LMAT mines that have a particular shape. Similarly, LMAT and HMAP objects of a particular shape share the second context. For classification results, we ran twenty experiments of 10-fold cross-validation. The average classification rates are given in Table 1. Each entry is described below.

- *Cl-HMM:* Sequences from each class are clustered into 4 using [15], and an HMM is learned for each cluster, resulting in 8 HMMs. A test sequence is assigned to the class whose HMM yields the highest log-likelihood.
- *Gate:* The first four HMMs are assumed to represent the first class, and the next four HMMs are assumed to represent the second class.
- *Experts:* Each expert HMM is used as a classifier.
- *MHMME + SVM:* For each sequence, the log-likelihoods obtained from all the HMMs in the MHMME model are concatenated to form a feature vector, which is then used in training an rbsvm with $\sigma = 1$.
- *PCA + ME:* The real and the imaginary parts of the data are combined to form a sequence of length 42. Then PCA is applied and the dimensionality is reduced to 10. These feature vectors are used to train a standard ME model.
- *PCA + SVM:* Similarly, the dimensionality is reduced to 10. An rbsvm is trained with $\sigma = 1$.
- *MCE-HMM:* Minimum Classification Error HMM is a discriminative learning method that minimizes the total misclassification error [7,16]. The parameters of MCE-HMM as they appear in [16] were set as follows: $\eta = 1$, $\gamma = 8$, $\theta = 0$, $\epsilon = 0.1$.

Table 1. 10-fold classification rates on landmine data

Model	Mean	Standard Deviation
MHMME + SVM	**0.83**	**0.04**
MHMME	0.80	0.05
PCA + SVM	0.78	0.04
MCE-HMM	0.75	0.05
PCA + ME	0.73	0.05
Gate	0.71	0.05
Cl-HMM	0.70	0.02
Experts	0.61	0.02

4. CONCLUSION

In this study simultaneous learning of context and classification has been addressed for sequential data, and the MHMME model has been developed. With landmine data experiments, it has been shown that the gates of MHMME partition the data in a way that each gate gives higher weights to certain types of sequences. Then, the experts classify these similar-looking sequences into mine and non-mine decisions. With these experiments, it has been shown that it is not just the gates or the experts, but it is the combination of both that results in good classification rates. In addition, MHMME significantly outperforms the HMM-based and ME-based algorithms.

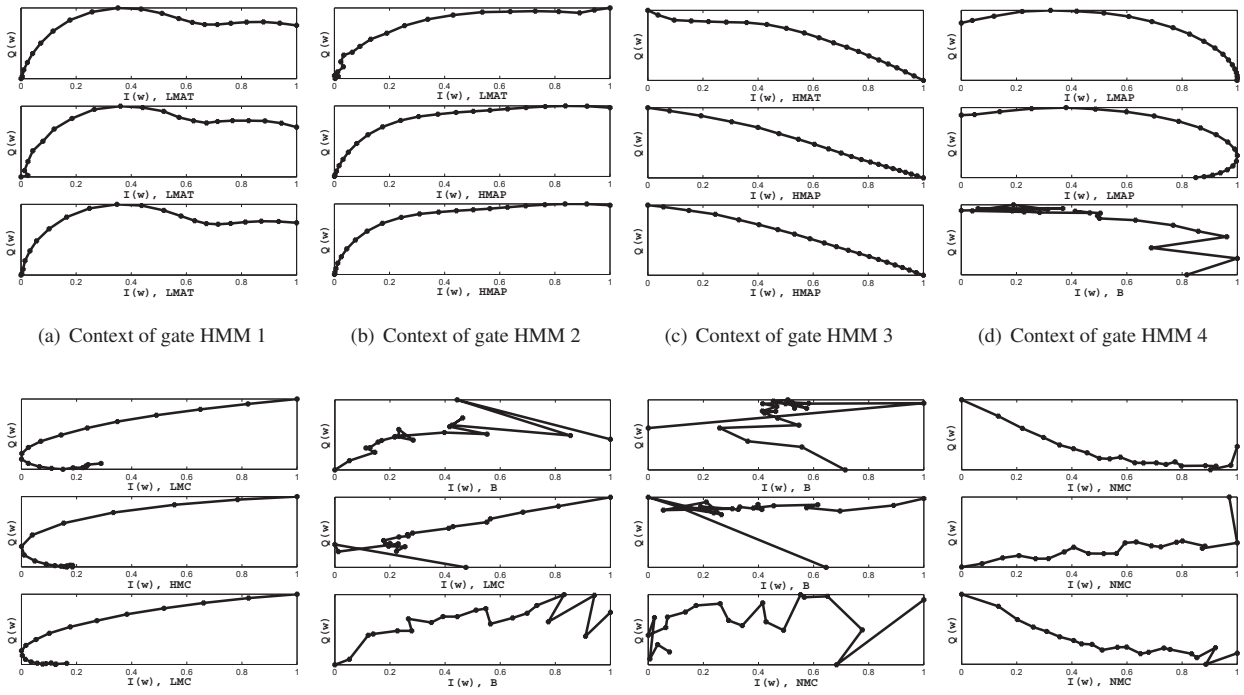
5. REFERENCES

[1] Michael I. Jordan, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181–214, 1994.

[2] Ke Chen, Dahong Xie, and Huisheng Chi, "A modified HME architecture for text-dependent speaker identification," *IEEE Transactions on Neural Networks*, vol. 7, pp. 1309–1313, 1996.

[3] Andreas S. Weigend, Morgan Mangeas, and Ashok N. Srivastava, "Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting," *International Journal of Neural Systems*, , no. 6, pp. 373–399, 1995.

[4] A.L.V. Coelho, C.A.M. Lima, and F.J. Von Zuben, "Hybrid genetic training of gated mixtures of experts for nonlinear time series forecasting," in *IEEE Int. Conf. on Systems, Man and Cybernetics*, 2003, vol. 5, pp. 4625–4630.



(a) Context of gate HMM 1 (b) Context of gate HMM 2 (c) Context of gate HMM 3 (d) Context of gate HMM 4

(e) Context of gate HMM 5 (f) Context of gate HMM 6 (g) Context of gate HMM 7 (h) Context of gate HMM 8

Fig. 2. Contexts defined by the gate HMMs. Each sequence is a normalized Argand diagram. Each column shows the top three Argand sequences that also represent the contexts. On the y-axis, the type of the mine or nonmine object is given.

[5] Xin Wang, Peter Whigham, Da Deng, and Martin Purvis, "Time-line hidden Markov experts for time series prediction," *Neural Information Processing - Letters and Reviews*, vol. 3, no. 2, pp. 39–48, May 2004.

[6] Zhiwu Lu, "A regularized minimum cross-entropy algorithm on mixtures of experts for time series prediction and curve detection," *Pattern Recognit. Lett.*, vol. 27, no. 9, pp. 947–955, 2006.

[7] Biing-Hwang Juang, Wu Hou, and Chin-Hui Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 3, pp. 257–265, May 1997.

[8] O. Missaoui, H. Frigui, and P. Gader, "Land-mine detection with ground-penetrating radar using multistream discrete hidden Markov models," *IEEE Trans. Geosci. Remote Sens.*, 2011.

[9] Eric B. Fails, Peter A. Torrione, Jr. Waymond R. Scott, and Leslie M. Collins, "Performance of a four parameter model for modeling landmine signatures in frequency domain wideband electromagnetic induction detection systems," in *SPIE Detection and Remediation Technologies for Mines and Minelike Targets XII*, 2007, pp. 65531–7.

[10] W.R. Scott, "Broadband array of electromagnetic induction sensors for detecting buried landmines," in *IEEE Int. Geoscience and Remote Sensing Symp.(IGARSS)*, July 2008, vol. 2, pp. 375–378.

[11] Seniha Esen Yuksel, Ganesan Ramachandran, Paul Gader, Joseph Wilson, Dominic Ho, and Gyeongyong Heo, "Hierarchical methods for landmine detection with wideband electromagnetic induction and ground penetrating radar multi-sensor systems," in *IEEE International Geoscience and Remote Sens. Symp. (IGARSS)*, July 2008, vol. 2, pp. II–177–II–180.

[12] G. Ramachandran, P.D. Gader, and J.N. Wilson, "GRANMA: Gradient angle model algorithm on wideband EMI data for land-mine detection," *IEEE Geosci. Remote Sens. Letters*, vol. 7, no. 3, pp. 535–539, July 2010.

[13] C. Ratto, P. Torrione, K. Morton, and L. Collins, "Context-dependent landmine detection with ground-penetrating radar using a hidden Markov context model," in *IEEE International Symposium on Geoscience and Remote Sensing (IGARSS)*, July 2010, pp. 4192–4195.

[14] H. Frigui, Lijun Zhang, and P.D. Gader, "Context-dependent multisensor fusion and its application to land mine detection," *IEEE Trans. Geoscience and Remote Sensing*, vol. 48, no. 6, pp. 2528–2543, June 2010.

[15] Padhraic Smyth, "Clustering sequences with hidden Markov models," in *Advances in Neural Inf. Proc. Systems (NIPS)*, 1997, pp. 648–654.

[16] Y. Zhao, P. Gader, P. Chen, and Y. Zhang, "Training dhmms of mine and clutter to minimize landmine detection errors," *IEEE Trans. Geosciences and Remote Sensing*, vol. 41, no. 5, pp. 1016–1024, May 2003.