34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Disease Progression Modeling Using Hidden Markov Models

Rafid Sukkar, *Senior Member, IEEE,* Elyse Katz, Yanwei Zhang, David Raunig, and Bradley T. Wyman

Abstract— **The development of novel treatments for many slowly progressing diseases, such as Alzheimer's disease (AD), is dependent on the ability to monitor and detect changes in disease progression. In some diseases the distinct clinical stages of the disease progress far too slowly to enable a quick evaluation of the efficacy of a given proposed treatment. To help improve the assessment of disease progression, we propose using Hidden Markov Models (HMM's) to model, in a more granular fashion, disease progression as compared to the clinical stages of the disease. Unlike many other applications of Hidden Markov Models, we train our HMM in an unsupervised way and then evaluate how effective the model is at uncovering underlying statistical patterns in disease progression by considering HMM states as disease stages. In this study, we focus on AD and show that our model, when evaluated on the cross validation data, can identify more granular disease stages than the three currently accepted clinical stages of "Normal", "MCI" (Mild Cognitive Impairment), and "AD".**

I. Introduction

The development of potential treatments for many slowly progressing diseases, such as Alzheimer's disease, can be helped by a disease progression model that can be used to readily detect disease progression or lack thereof. This implies that the model should be able to detect more granular stages in the disease as compared to disease stages corresponding to clinical diagnoses. Several research efforts have focused on disease progression modeling and prediction [1-3]. In [1] regression analysis is used on a set of clinical measurements to assess and predict disease progression, while in [2] the authors propose a non-linear model based on the longitudinal change of Alzheimer's disease Assessment Scale Scores. An event-based model for disease progression is introduced in [3] where disease progression is modeled as a series of events defined as significant changes in symptoms or in tissue. In this work we employ a set of biomarkers in a data-driven statistical framework based on Hidden Markov Models (HMM's) to model disease progression.

HMM's have been successfully used in many areas to model and classify sequences and time signals. For example, HMM's have been extensively used in speech recognition [4], genome analysis [5] and handwriting recognition [6]. Here we propose using HMM's with a set of biomarkers as the HMM features to model disease progression with a focus of AD. Unlike the applications mentioned above, our aim is not to detect and classify a given sequence, but rather to uncover more granular disease stages as compared with stages corresponding to clinical diagnoses. As such, our HMM training strategy differs from training for classification purposes where supervised training is carried out resulting in different HMM's for different classes. Here the classes (i.e., disease stages) that we are trying to uncover are unknown or latent. Accordingly, we perform HMM training in an unsupervised way to allow the model to exploit temporal statistical patterns in the biomarker signal to freely cluster together or pull apart different stages of the disease based on statistical similarity of the data within each cluster. We base our HMM training and cross validation on a database of longitudinal biomarker measurements of subjects with different stages of the disease. Once the model is trained, we interpret each state in the model as a stage in disease progression. We validate this interpretation by data-driven evaluation of the trained model on the cross validation data

The paper is organized as follows. In the next section a brief description of Hidden Markov Models is given and in Section III, the proposed model for disease progression is introduced. This is followed in Section IV by a description of the HMM model for AD progression. In Section V, we present experimental results, followed by conclusions.

II. Hidden Markov Models

A Hidden Markov Model consists of a set of interconnected states where the connections are governed by a set of transitional probabilities. What sets a Hidden Markov Model apart from a Markov Model or a Markov Chain is the fact that in a Markov Chain the states are observable while for Hidden Markov Models, the states are statistical having associated probability distributions called the observation probability density functions. The observation is typically a multidimensional vector consisting of a set of features called the HMM feature vector. The observation density functions can either be continuous or discrete. In this work we will be using continuous distributions based on Gaussian Mixtures [4]. Maximum likelihood training using the Expectation-Maximization (E-M) iterative algorithm is commonly used to estimate the HMM parameters [4].

III. Modeling Disease Progression

We consider the modeling of disease progression as the problem of modeling the evolution of a set of biomarker features in time where the choice of biomarkers is motivated

Rafid Sukkar is with Voxelon, Inc., Niles, IL 60714 USA (e-mail: rafids@voxelon.com)

Elyse Katz was with Pfizer, Inc., Groton, CT 06340 USA (e-mail: ekatz222@gmail.com).

Yanwei Zhang is with Pfizer, Inc., Groton, CT 06340 USA (e-mail: Yanwei.Zhang@pfizer.com).

David Raunig was with Pfizer, Inc., Groton, CT 06340 USA. He is now with ICON Medical Imaging, Warrington, PA 18976 USA (email: David.Raunig@iconplc.com)

Bradley T. Wyman is with Pfizer, Inc., Groton, CT 06340 USA (email: Brad.Wyman@pfizer.com)

by their ability to discriminate among the disease stages. The goal is to model the disease progression in a more granular fashion as compared with the known clinical disease stages.

Figure 1 shows the topology of the Hidden Markov Model that we propose for modeling disease progression. In this figure a_{ij} is the transitional probability from state i to state j, $b_i(\boldsymbol{x})$ is observation probability density function of state i, $\boldsymbol{x}$ is the HMM feature vector, and c_i is the *a-priori* probability of starting in state i. Since disease progresses with time, we propose a left-to-right topology, as Figure 1 shows, where we will consider each HMM state as a disease stage. A transition that occurs from one state to the next indicates disease progression while a transition to the same state indicates no progression. In the topology of Figure 1, we also include a transition to a previous state. Depending on the specific disease that we are modeling, such a transition can indicate disease regression due to, for instance, the presence of a potential treatment or because the disease regressed naturally. It is also included to model rare inaccuracies in the measurements of the biomarkers used in the HMM feature vector. Although figure 1 shows a 6-state HMM, any number of states can be used depending on the specific disease, the number of desired disease stages, and the size of the available HMM training data.

IV. Alzheimer's Disease Progression Model

The model of Figure 1 with 6 HMM states is used in this work for Alzheimer's disease progression model. As mentioned above, we will consider each state to be a disease stage implying that the model will result in 6 modeled stages of the disease as compared to the three currently defined clinical stages of "Normal", "MCI" (Mild Cognitive Impairment), and "AD" (Alzheimer's Disease).

Typically HMM's are used in the classification of competing classes where a separate HMM is trained to model the evolution of the signal for a given class. Subsequently, in the testing phase, an unknown signal is evaluated against all competing class HMMs and the one with the highest score is chosen as the classification answer. In this work, we train and test the HMM in a slightly different approach. Since our goal is to uncover and model stages of the disease, the HMM training is carried out in an unsupervised way where the time signals from all subjects, irrespective of their clinical diagnosis at any given time, are used in the training process. The idea here is to let the training strategy of the HMM exploit patterns in the biomarker feature vector both temporally and across the individual biomarker features to cluster similar conditions of the subjects together in a given state and to separate different conditions into different states. Since disease progresses in time, subsequent states will indicate further progression of the disease.

The data used in our modeling is the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set [7]. The ADNI data includes longitudinal biomarker measurements of 819 subjects. At the start of the longitudinal study 229 were clinically diagnosed as "Normal", 398 were "MCI", and 192 were "AD". The subjects were followed for a period of up to 36 months with periodic visits every 6 months when clinical evaluations were performed and biomarker measurements taken. As time progressed the clinical diagnosis of some of the subjects progressed while others remained the same.

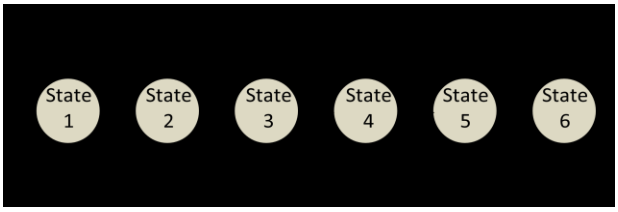

Figure 1. HMM topology and parameters for disease progression modeling.

It has been shown that the brain ventricular volume and hippocampus volume as measured by MRI are correlated with AD diagnosis [8,9]. Using the ADNI dataset, we show in Figure 2, a scatter plot of the hippocampus volume normalized by the skull volume versus the Ventricular Boundary Shift Integral (VBSI) biomarker for Normal and AD diagnoses. The VBSI is a measure of the change in brain ventricular volume from a baseline [10]. It is clear from Figure 2 that although there is some overlap, the majority of the data show a good degree of separation between the two classes. Based on these observations, we form the HMM biomarker feature vector as a 4-dimenstional vector consisting of the above two biomarkers along with their rate of change over time. The incorporation of the dynamic rate-of-change features in the HMM feature vector have been successfully used in the context of speech recognition [4]. In this work, we represent the rate of change as the change in biomarker value between two successive visits. Specifically, the HMM feature vector used here consists of the following parameters:

1. Ventricular Boundary Shift Integral (VBSI)
2. Hippocampus volume normalized by the skull volume,
3. Change in VBSI between two successive visits
4. Change in normalized hippocampus volume between two successive visits

The ADNI dataset includes a confidence rating for the biomarkers of our feature vector. This confidence rating is assigned by the organization or lab that took the biomarker measurement. In our experiments, we excluded biomarker measurements that have low confidence. In addition, since our HMM feature vector includes dynamic features, subjects with less than 2 visits worth of biomarker measurements in the ADNI data were excluded. The above criteria resulted in data from 594 subjects used in our experiments.

V. Experimental Results

Data from the 594 ADNI subjects were randomly partitioned into a training set and a testing (cross validation) set. We used 70% of the subjects (416) for training and the remaining subjects for cross validation. Based on the topology of Figure 1, HMM training was performed in an unsupervised way where any subject regardless of the clinical diagnosis at any of his/her visits was allowed to enter the HMM at any state and progress through the model, and then end at any state. In this fashion, the training process is given the freedom to cluster data points with similar stages of the disease together into one state guided by the topology of Figure 1.

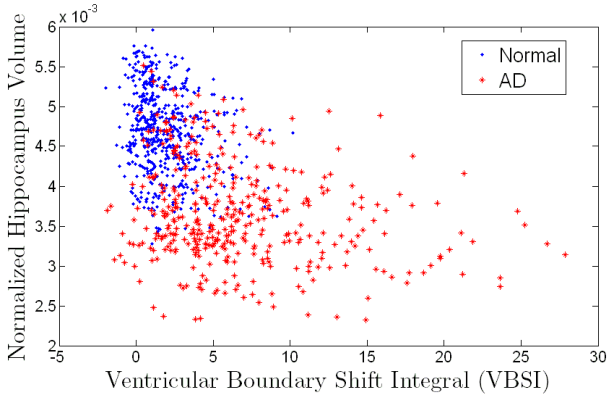

Figure 2. Scatter plot of the normalized hippocampus volume versus VBSI for Normal and AD subjects.

Cross validation is performed by processing the testing set with the trained HMM. The Viterbi algorithm is used to determine the optimal maximum likelihood state sequence for each subject given his/her biomarker measurements at each visit. To evaluate the model for its performance in modeling disease progression, the HMM state sequences were correlated with the subjects' actual clinical diagnosis at the corresponding visit. As a result, we have, for each subject, an HMM state sequence and a corresponding sequence of clinical diagnoses. Using this data, we then compute the probability of a specific diagnosis given an HMM state, P[Diagnosis|State]. Figure 3 shows two plots of these probabilities for the three classes of diagnoses given the HMM states computed over the testing and training set, respectively. We can see that "Normal" diagnosis dominates in early (i.e., low index) states and diminishes with increasing state index, while "AD" diagnosis behaves in the opposite way monotonically increasing with state index. The "MCI" diagnosis, dominates in states 4 and 3 for the testing and training set, respectively, and diminishes as we move away from these middle states. Given that the clinical diagnoses are progressing stages of the disease, we can interpret these results as evidence that the states of the HMM represent varying and more granular stages in disease progression. To gain more insight into the performance of the HMM in modeling disease progression, we used a measurement called the Clinical Dementia Rating Scale Sum of Boxes (CDR-SB). The CDR-SB score is derived from patient interviews and mental status examination and is included in the ADNI dataset for each visit. The score ranges from 0-8 where higher scores indicate higher dementia impairment and correlates with Alzheimer's disease progression [11]. To relate the CDR-SB score to the trained HMM, we determined a CDR-SB value for each state in the model. We did so, by processing the training set past the model and determining the optimal state sequence for all the subjects in that set. Then, we computed the average CDR-SB score over all visits that dwelled in a given HMM state given the optimal state sequence for each subject. Specifically, we compute, s_i, the CDR-SB score for state i over the training set, as follows:

$$s_i = \frac{1}{D_i} \sum_{k=1}^{K} \sum_{n=1}^{N_k} b_{k,n}\, \delta_{v(n,k),i} \qquad 1 \le i \le 6 \tag{1}$$

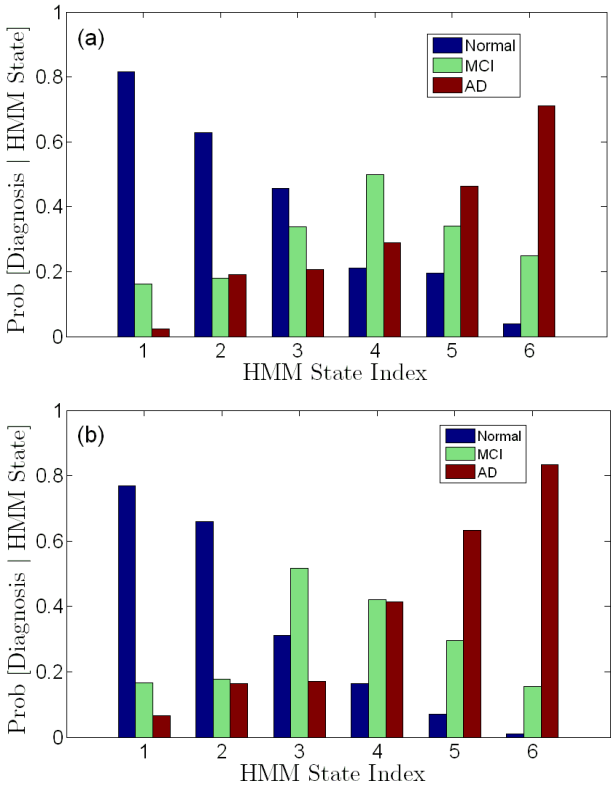

Figure 3. Probability of clinical diagnosis over all visits that the optimal state sequence assigned to a given state, (a) Testing set (b) Training set.

where K is the number of subjects in the training set, N_k is the number of visits for subject k, $b_{k,n}$ is the CDR-SB score for subject k at visit n, the function $v(k,n)$ maps subject $k's$ n^{th} visit to the state index as dictated by the optimal HMM state sequence, $\delta_{j,k}$ is the Kronecker delta, and $D_i = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \delta_{v(n,k),i}$. Figure 4 shows two plots of s_i, one computed over the training set as Equation (1) indicates, and the other computed over the testing set. This figure clearly shows that for both the training and testing sets there is a monotonic relationship between the state CDR-SB score and the state index indicating that the states of the model correlate with disease progression.

To see how each individual subject progresses through the model relative to the CDR-SB scores of the subject's visit, we compute the root mean square deviation between the sequence of s_i values along the optimal state sequence for a given testing set subject and the actual CDR-SB score of each visit for the subject, as follows:

$$e_l = \left[\frac{1}{N_l} \sum_{n=1}^{N_l} \left(s_{v(l,n)} - b_{l,n} \right)^2 \right]^{1/2} \qquad 1 \le l \le L \tag{2}$$

where L is the total number of subjects in the testing set, N_l is the number of visits for subject l, and e_l is the root mean squared deviation for subject l. Figure 5 shows a histogram of e_l computed across all subjects in the testing set. Although the majority of the subjects exhibit low deviations, 25% of the subjects have deviations greater than 2. This suggests that

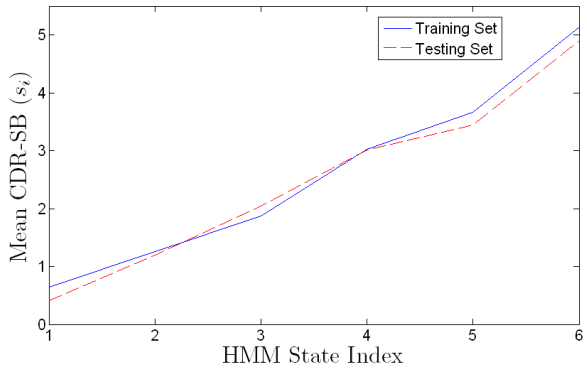

Figure 4. Mean CDR-SB computed over all visits that dwelled in a given HMM state versus HMM state index.

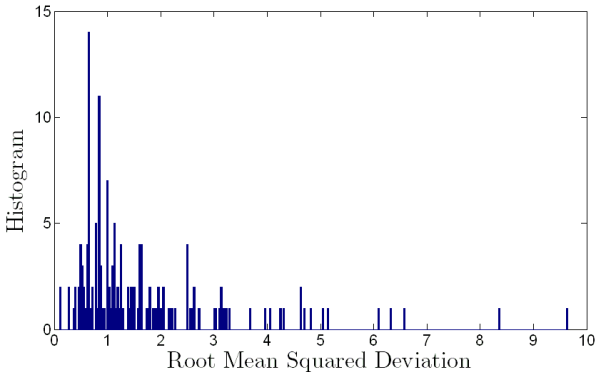

Figure 5. Histogram of subject CDR-SB Root Mean Squared Deviation, e_l, computed over the testing set.

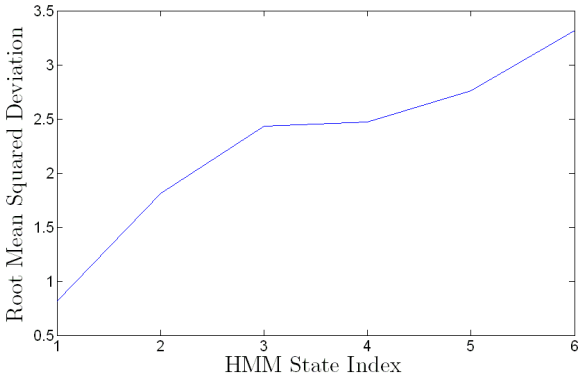

Figure 6. HMM state CDR-SB Root Mean Squared Deviation computed over all test set subjects visits that dwelled in a given HMM state versus HMM state index.

while the results of Figure 4 give evidence of disease progression as a function of state index with a semi-linear relationship between state indexes and the mean CDR-SB score, s_i, Figure 5 results show that, for some subjects in the cross validation set, the HMM provides a different disease progression path than that indicated by the CDR-SB scores. Such a different path may provide more revealing aspects of the progression of the disease. To gain more insight into how the progression paths of the HMM and CDR-SB compare, Figure 6 shows the Root Mean Squared Deviation computed across all testing set subjects visits that dwelled in a particular HMM state according to the optimal HMM state sequence. We see here that the deviation at the low HMM state indexes are lower than at the high state indexes. This suggests that at the normal and early stages of the disease the HMM and the CDR-SB indicate similar progression path. However, as the disease progresses, the HMM provides an increasingly different disease progression path which can give a different perspective on the progression of the disease.

VI. Conclusions

We presented a model for disease progression based on a Hidden Markov Model framework. Using the ADNI data set biomarkers for Alzheimer's disease, we trained an HMM in an unsupervised way with the goal of uncovering more granular stages in disease progression. We showed that the trained HMM is able to model disease progression more granularly than the currently defined clinical stages.

References

[1] R. S. Doody, V. Pavlik, P. Massman, S. Routree, E. Darby, and W. Chan, "Predicting Progression of Alzheimer's Disease,", *Alzheimer's Research and Therapy*, Feb. 2010.

[2] M. N. Samtani, M. Farnum, V.Lobanov, E. Yang, N. Raghavan, A. DiBernardo, and V. Narayan, "An Improved Model for Disease Progression in Patients from the Alzheimer's Disease Neuroimaging Initiative," *J. Clinical Pharmacology,* June 2009.

[3] H. M. Fonteijn, M. J. Clarkson, M. Modat, J. Barnes, M. Lehmann, S. Ourselin, N. C. Fox, and D. C. Alexander, "An Event-Based Disease Progression Model and Its Application to Familial Alzheimer's Disease," *Lecture Notes in Computer Science*, Vol. 6801/2011, pp. 748-759, 2011

[4] X. Huang, A. Acero, H.W. Hon, *Spoken Language Processing,* Prentice Hall PTR, 2001.

[5] S.R. Eddy, "Hidden Markov Models and Large-Scale Genome Analysis," *Trans. American Crystallographic Assoc.,* 1997.

[6] H. Jianying, M. K. Brown, and W. Turin, "HMM Based Online Handwriting Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 18, pp. 1039-1045, Oct. 1996.

[7] http://www.adni-info.org

[8] S. M. Nestor, R. Rupsingh, M. Borrie, M. Smith, V. Accomazzi, J. L. Wells, J. Fogarty, and R. Bartha, "Ventricular Enlargement as a Possible Measure of Alzheimer's Disease Progression Validated Using the Alzheimer's Disease Neuroimaging Initiative Database," *Brain: a Journal of Neurology,* pp. 2443-2354, Sept. 2008.

[9] P. Scheltens, D. Leys, F. Barkhof, D. Huglo, H.C. Weinstein, P. Vermersch, M. Kuiper, M. Steinling, E. C. Wolters, and J. Valk, "Atrophy of Medial Temporal Lobes on MRI in "Probable" Alzheimer's Disease and Normal Ageing: Diagnostic Value and Neuropsychological Correlates," *Journal of Neurology, Neurosurgery, and Psychiatry*, pp. 967-972, 1992.

[10] E. Gordon , J. D. Rohrer , L. G. Kim L, R. Omar , M. N. Rossor, N. C. Fox, J. D. Warren, "Measuring Disease Progression in Frontotemporal Lobar Degeneration: a Clinical and MRI Study," *Neurology,* pp. 666-673, Feb 2010.

[11] S. E. O'Bryant, S. C. Waring, C. M. Cullum, J. Hall, L. Lacritz, P. J. Massman, P. J. Lupo, J. S. Reisch, R. Doody, "Staging dementia using Clinical Dementia Rating Scale Sum of Boxes scores," *Archives of Neurology*, pp. 1091-1095, Aug. 2008.