View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Columbia University Academic Commons

Bayesian Transduction and Markov Conditional Mixtures for Spatiotemporal Interactive Segmentation

Noah Lee and Andrew F. Laine
Department of Biomedical Engineering
Heffner Biomedical Imaging Lab
Columbia University
New York, NY USA
{nl2168, laine}@columbia.edu

Shahram Ebadollahi
T. J. Watson Research Center
IBM Research
Hawthorne, NY USA
ebad@us.ibm.com

Robert L. DeLaPaz
Department of Neuroradiology
Columbia University College of Physicians & Surgeons
New York, NY USA
rld17@columbia.edu

***Abstract* - In this paper we propose a novel transductive learning machine for spatiotemporal classification casted as an interactive segmentation problem. We present Markov conditional mixtures of naïve Bayes models with spatiotemporal regularization constraints in a transductive learning and inference framework. The proposed model extends on previous work [3] to account for non independent and identically distributed (i.i.d.) sequential data by imposing the learning and inference problem w.r.t. time. The multimodal mixture assumption on the class-conditional likelihood for each covariate feature domain in conjunction with spatiotemporal regularization constraints allow us to explain more complex distributions required for classification in multimodal longitudinal brain imagery. We evaluate the proposed algorithm on multimodal temporal MRI brain images using ROC statistics and report preliminary results.**

Neural Informatics, Spatiotemporal Interactive Segmentation, Naïve Bayesian Transduction, Markov Conditional Mixtures.

I. INTRODUCTION

Over the last decade we have seen an increase in the amount and complexity of heterogeneous biomedical data coming from laboratory tests, imaging methods, and gene-protein analysis revealing information at nano-to-organ scales. Current clinical practice of neuro-oncology requires the physician to manually extract, correlate and interpret information from heterogeneous data sources over time to diagnose, treat and manage patients with brain tumors [4]. Multimodal brain imaging provides the physician with neuro-pathologic and -anatomic information enabling quantitative assessment of tumor progression and side effect symptoms. In the medical domain the labeling process of multimodal spatiotemporal image data requires expert knowledge and time intensive editing to obtain accurate label information for the object that is to be quantified. In the realm of computer aided diagnosis (CAD) interactive segmentation schemes have been well received by physicians, where the combination of human and machine intelligence can provide improved segmentation efficacy with minimal expert intervention [5]. Transductive learning (TL) or semi-supervised learning (SSL) is a suitable framework for learning-based interactive spatiotemporal segmentation given the scarce label problem. In this regard, transduction offers a workaround by leveraging the labels provided at time t to label the remaining test set at time $t+1$. Obtaining temporal label information of the object of interest enables quantification of neuro-pathologic information that can be used for exploring functional and temporal relationships to other indicative factors to assess the patient's health condition over time [4].

In this paper we propose a novel transductive learning machine for spatiotemporal classification casted as an interactive segmentation problem. The contribution of our paper lies in extension of previous work [3] to the spatiotemporal domain. We present Markov conditional mixtures of naïve Bayes models (T-MCMNB) with spatiotemporal regularization constraints in a transductive learning and inference framework. The transductive generative formalism w.r.t. time allows us to provide i) predictive confidence of the classification for non i.i.d. sequential data and ii) assess performance guarantees of the inference while exploiting correlation between temporal observations. In a probabilistic formulation and using the framework of graphical models [8] we consider a bounded probability measure $\mathcal{P}_{XY}$ describing the joint distribution of the given input and output label space $X \times \mathbb{Z}^K$. We make use of unconditional and conditional regularized Gaussian mixture models for each covariate feature domain on the class-conditional likelihood to learn and infer the relationships in $\mathcal{P}_{XY}$ using naïve Bayesian transduction. The naïve conditional independence assumption allows efficient computation of marginal and conditional distributions [6] for large-scale learning and inference. We choose a generative model over the discriminative counterpart motivated in part by a faster convergence rate of the asymptotic generalization error [9] when label information is scarce. Since the goal is to obtain label information only for the unlabeled test set at time $t+1$ we allow the posterior distribution to depend on future observations with spatiotemporal regularization constraints to exploit the smoothness- and cluster assumption between $\mathcal{P}_{X\,t,t+1}$ and $\mathcal{P}(y|x_{t,t+1})$. The algorithm shows promising segmentation performance with a sensitivity and specificity of up to 93.16% and 99.91%.

II. RELATED WORK

Consider a dataset $\mathcal{D} = [\mathcal{D}_l, \mathcal{D}_u]$drawn non-i.i.d. from $\mathcal{P}_{XY}$, where $\mathcal{D}_l = \{(\mathbf{x}_n^t, y_n^t)\}_{n=1}^{l}$ denote the labeled training set and $\mathcal{D}_u = \{(\mathbf{x}_n^{t+1}, \hat{y}_n^{t+1})\}_{n=l+1}^{l+u}$ the unlabeled test set with $\hat{y}_n^{t+1}$ unknown. The usual case is, that $l << u$.

A. The Naïve Bayes Model

\ The naïve Bayes classifier finds successful application in text categorization tasks [2, 7]. Let $\mathcal{P}(y, x_1, ..., x_n)$ denote the joint distribution of the input samples and the class labels. The naïve conditional independence assumption allows us to factorize the joint distribution as a product of class prior and independent conditional probability distributions $\mathcal{P}(y)\prod_{n=1}^{N}\mathcal{P}(x_n|y)$. In graphical model notation the naïve Bayes model has for each X_j node the parent node Y, where j indexes the covariate feature dimension and n the number of samples. For the discrete case we assume each X_j to be sampled from a multinomial probability model $p(x) = \prod_{m=1}^{M}\vec{\alpha}(m)^{\vec{x}(m)}$ with $\sum_m \vec{\alpha}(m) = 1$. The class-conditional probability for each X_j for the continuous case takes the form of a Gaussian $p(x) \sim \mathcal{N}(\mu_j|\sigma_j)$ with $\mu_{jk} = E[X_j|Y = y_k]$ and $\sigma_{jk}^2 = E[(X_j - \mu_{jk})^2|Y = y_k]$.

B. Transductive Multinomial Naïve Bayes

The transductive naïve Bayes classifier [1, 2] was introduced for the application of text classification. The classifier uses both the training documents and the distribution of the test documents to learn a classification rule. The model is similar to the one outlined in section 2.A with the extension to perform transductive inference. The algorithm classifies the test documents using a multinomial naïve Bayes model initially learned from the labeled training documents (Step-I) and then sequentially relearned on the classified unlabeled test documents (Step-II) to perform transduction. We summarize their model as follows by omitting the time index:

$$\begin{gathered}
\textbf{Step-I} \\
\mathcal{L}(\hat{\theta}|\mathcal{D}_l) = \sum_{n=1}^{l}\log\, p(y_n|\pi) + \sum_{n=1}^{l}\sum_{j=1}^{d}\log\, p(x_{j,n}|y_n, \phi) \\
\textbf{Step-II} \\
\mathcal{L}(\theta|\mathcal{D}_{l+u}) = \sum_{n=l+1}^{l+u}\log\, \hat{p}(y_n|\hat{\pi}) + \sum_{n=l+1}^{l+u}\sum_{j=1}^{d}\log\, \hat{p}(x_{j,n}|y_n, \hat{\phi}) \\
\text{subject to } \sum_m \hat{\phi}_{kjm} = 1.
\end{gathered} \tag{1}$$

This two-step iterative scheme estimates the prior and the class-conditional probability of the naïve Bayes model taking into account the unlabeled test distribution. Here $\hat{\theta}$ denotes the maximum a posteriori (MAP) estimate obtained from the labeled training set. They propagate into Step-II indicating that they have been iteratively relearned on the classified unlabeled test set. As reviewed in section 2.A the model in (1) assumes a multinomial probability model on the data when computing the class-conditional likelihood making it non-applicable to multimodal continuous data domains such as in multimodal medical brain imagery.

Exponential family models such as Gaussian or multinomial class-conditional mixture models may be restrictive dependent on the modeling problem and application domain. In real world applications often times the single Gaussian assumption is to limited to fully explain the complexity of $\mathcal{P}_X$. In non-negative data domains the uniform Gaussian assumption may produce incorrect model behavior due to variance symmetry or insufficient descriptive power. Previously outlined multinomial probability models mainly used in text classification [7] or their multinomial mixture counterpart [2] assume discrete finite unordered data domains with a fixed set of values. As opposed to [1, 2] we allow $\mathcal{P}_X$ to be continuous and non-uniformly distributed with a multimodal cluster and smoothness assumption. Moreover we transfer this assumption into the temporal domain by letting $\mathcal{D}_u$ come from $\mathcal{P}_{X^{t+1}Y}$and $\mathcal{D}_l$ from $\mathcal{P}_{X^tY}$.

III. TRANSDUCTIVE MARKOV CONDITIONAL MIXTURE NAÏVE BAYES FOR SPATIOTEMPORAL SEGMENTATION

Given the scarce label problem we choose a generative model over the discriminative counterpart motivated in part by a faster convergence rate of the asymptotic generalization error [9] when label information is scarce. In particular, we present a Markov conditional mixture naïve Bayes model (T-MCMNB) with spatial regularization constraints in a transductive learning and inference setting. Compared to [1] and [2] our model assumes for the class-conditional likelihood a spherical Gaussian mixture allowing us to represent and describe more complex distributions while keeping the parameter space tractable. To simplify the estimation we reduce the parameter space by assuming naïve conditional independence between the feature space and the class label imposing a regularization constraint on the class-conditional likelihood. The naïve conditional independence assumption allows efficient computation of marginal and conditional distributions [6] suitable for large scale learning and inference. The posterior is formed by learning the class-conditional mixture model $p(x|y)$ and prior $p(y)$ for each class exploiting labeled and unlabeled data. We allow the posterior distribution to depend on the unlabeled test set $\mathcal{D}_u$ with spatial regularization constraints to exploit the smoothness- and cluster assumption between $\mathcal{P}_{X\,t,t+1}$ and $\mathcal{P}(y|x_{t,t+1})$ to perform predictive temporal segmentation. Transductive learning and inference assume and exploit a cluster assumption, where each cluster reflects different distributions of different species or latent model classes.

A. Conditional Multi Latent Variable Model

Our modeling problem consists of two latent variables one for $\mathcal{P}(y|x_{t,t+1})$ and the other for approximating the marginal $\mathcal{P}_{X_{t,t+1}}$. To account for multimodal densities we can consider a sub probability model $f_c(x_j|\theta_c)$ for each component $c \in C$. One can build an unconditional *mixture density* on $\mathcal{P}_{X_{t,t+1}}$ with

$$p(x_j|\theta) = \sum_{c=1}^{C} p(x_j, z^c = 1|\theta) = \sum_{c=1}^{C} \alpha_c f_c(x_j|\theta_c), \quad \text{subject to } \sum_{c=1}^{C} \alpha_c = 1, \alpha \geq 0, \tag{2}$$

where $f_c(x|\theta_c)$ are the *mixture components* obtained by marginalizing and conditioning over a latent or hidden variable Z. The non-negativity constraints α_c are the *mixing proportions* and $\theta \equiv (\alpha_c, \theta_c)_{c=1}^{C}$ denote the parameter space. In generative graphical models the latent variable forms the parent over the data leading us to the problem of density estimation. Rather than estimating an unknown density in our case we are interested in inferring class labels with observed latent variables using a conditional mixture model on the data. Using Bayes rule one can achieve this task by inverting the *mixture density* model in (2) to perform probabilistic inference. Conditioning on $\mathcal{P}_{X_{t,t+1}}$ the class-conditional of the latent variable Y is

$$p(y^k = 1|\mathbf{x}^t, \theta) = \frac{\alpha_{k,c} f_{k,c}(\mathbf{x}^t|\theta_{k,c})}{\sum_j \alpha_j f_j(\mathbf{x}^t|\theta_j)} \tag{3}$$

The knowledge of $\mathbf{x}^t$ and θ enables us to obtain the probability of the unobserved latent class label Y given the data. From the learned probability model at time t we can predict the class distribution of the next time step $t+1$ to perform temporal segmentation.

B. Transductive Learning and Inference

Given $\mathcal{D}_l$ we learn the class-conditional and unconditional mixture densities of each class by maximizing the log-likelihood of $p(x|y)$ and $p(y)$. To learn the marginal $\mathcal{P}_X$ for a given class label we assume $p(x)$ to be distributed as a spherical Gaussian mixture. To approximate both latent variables Z and Y we build the following likelihood model on $\mathcal{D}_l$ and $\mathcal{D}_u$

$$\mathcal{P}(\mathcal{D}|\theta) = \prod_{n=l+1}^{l} p(y_n|\pi) \sum_{j=1}^{d} \alpha_j \; p(x_{jn}^t|y_n, \theta) \times \prod_{n=l+1}^{l+u} p(y_n|\pi) \sum_{j=1}^{d} \alpha_j \; p(x_{jn}^{t+1}|y_n, \theta). \tag{4}$$

The MAP estimate of parameter θ for $\mathcal{D} = [\mathcal{D}_l, \mathcal{D}_u] \in \mathcal{P}_{XY}$ with $l+u$ i.i.d. observations has no closed form solution. Taking the log-likelihood of (4) gives

$$\mathcal{L}(\hat{\theta}|\mathcal{D}) = \sum_{n=1}^{l} \log \, p(y_n|\pi) + \sum_{n=1}^{l} \log \sum_{j=1}^{d} \alpha_j p(x_{jn}^t|y_n, \theta) + \sum_{n=l+1}^{l+u} \log \, \hat{p}(y_n|\hat{\pi}) + \sum_{n=l+1}^{l+u} \log \sum_{j=1}^{d} \alpha_j \hat{p}(x_{jn}^{t+1}|y_n, \hat{\theta}) \tag{5}$$

The log-sum term of above log-likelihood in equation (5) is a marginal probability and requires a non-linear optimization scheme. Alternatively equation (5) can be optimized by an iterative method to obtain the MAP or ML solution. One can choose from belief propagation and other approximate inference algorithms in probabilistic graphical models [8]. We choose the EM algorithm [10] for the sake of simplicity and conceptual clarity. Lower bounding the log-sum term with an auxiliary function $\mathcal{L}(q,\theta)$ a local solution can be obtained by iteratively ascending

$$\begin{aligned} &\textbf{E-Step} \quad q^{t+1} = \arg\max_{q} \mathcal{L}(q, \theta^{(i)}) \\ &\textbf{M-Step} \quad \theta^{(i+1)} = \arg\max_{\theta} \mathcal{L}(q^{(i+1)}, \theta). \end{aligned} \tag{6}$$

The first term of equation (6) calculates the posterior probability (E-step) whereas the preceding steps in (5) are the (M-step) equations. A proof that the update equations in (6) indeed maximize the log-likelihood can be found in [10].

The ML estimate of the sum-log term of equation (5) is much simpler. Maximizing the log-likelihood with respect to π the solution to the constraint optimization problem for the labeled training data is:

$$\hat{\pi}_{ML} = \arg\max_{\pi} \sum_{n=1}^{l} \log \, p(y_n|\pi) = \sum_{n=1}^{l} y_n / l \tag{7}$$

Analog to equation (7) the ML estimate for π on the unlabeled test set updates accordingly with changed summation indices. From (6) and (7) the maximum a posteriori classification on the unlabeled test set can be obtained in a straight forward manner.

IV. Results & Discussion

We applied our algorithm to the task of interactive spatiotemporal brain tumor (edema) segmentation and evaluated our method with quantitative comparison to expert grading. We performed experimental evaluation on a multimodal temporal MR medical brain dataset with $X_t, X_{t+1} \in \mathcal{R}^{D=2}$ and $Y \in \mathbb{Z}^{K=2}$. The dataset has a resolution of 256x256x30 per modality and anisotropic voxel dimensions of 0.4x0.4x5mm. Multimodal registration was applied to bring the multimodal data sources into a common coordinate frame. We selected the most simplistic configuration of multimodal features by looking at FLAIR and DWI voxel intensities. The algorithm showed promising segmentation performance with a sensitivity and specificity of up to 93.16% and 99.91% respectively. By labeling a single slice at time t accurate predictions for future observations could be obtained.

V. Conclusion

In this paper we have presented a novel algorithm for spatiotemporal interactive segmentation using Bayesian transduction and Markov conditional mixtures in naïve Bayes models. By imposing the transductive learning and inference problem w.r.t. time and in conjunction with spatiotemporal regularization constraints efficient segmentation in non i.i.d. data could be achieved. Future work is devoted to extend the algorithm to higher-order Markov constraints to enable temporal active learning for interactive segmentation.

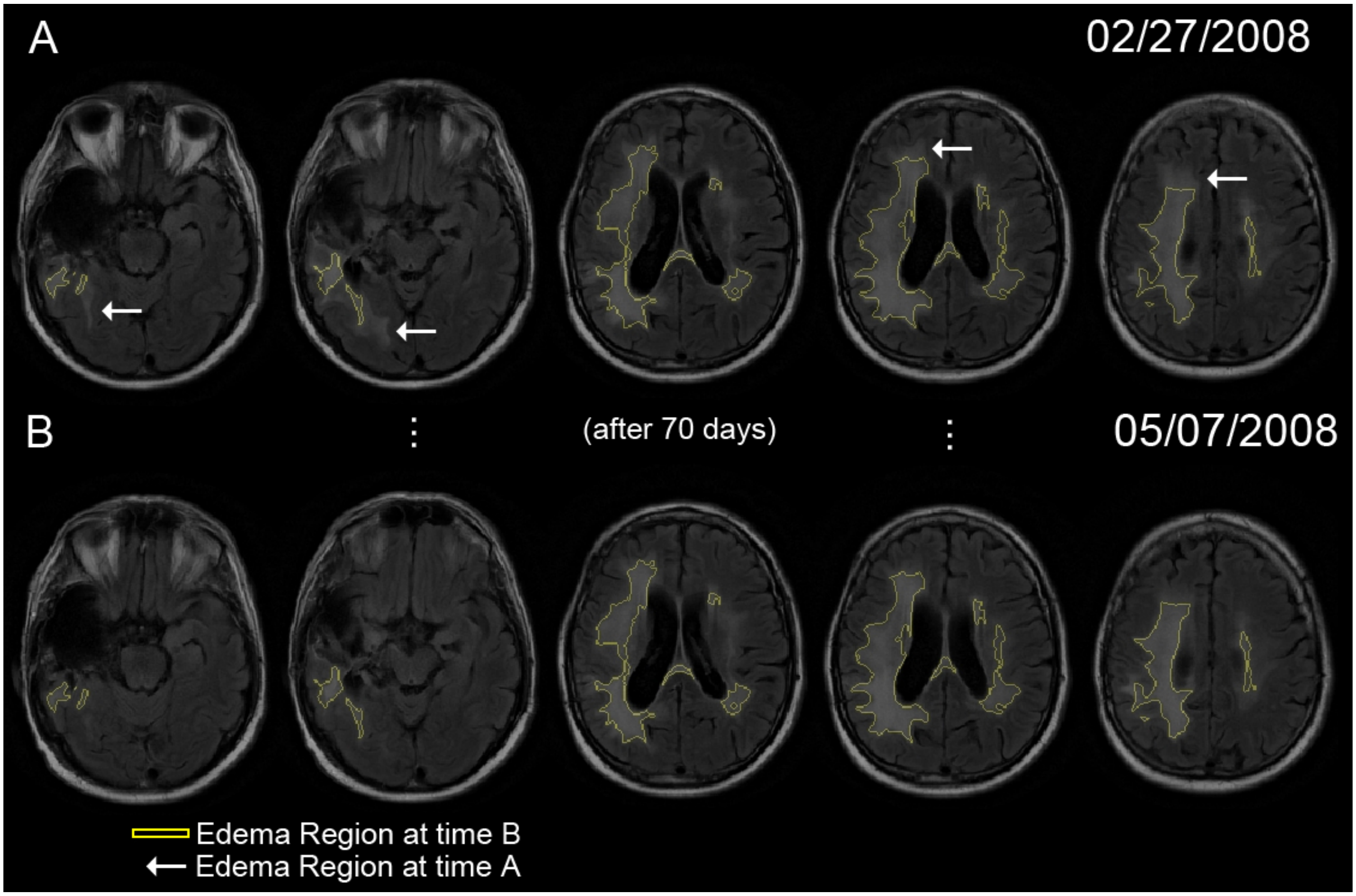

Figure 1. Qualitative spatiotemporal segmentation results using T- MCMNB. A) Segmentation contours predicted for B are overlaid on dataset A to show edema change. The white arrows point out edema regions at time A that disappear in later. B) Segmentation result predicted for dataset B showing partial edema shrinkage. From left to right different slices of the brain are shown.

References

[1] Branson K., "A Naïve Bayes Classifier Using Transductive Inference for Text Classification", Technical Report, Dept. of Computer Science and Engineering, UCSD (2001). Unpublished.

[2] Nigam K., McCallum A., Thrun S., and Mitchell T., "Text Classification from Labeled and Unlabeled Documents using EM", Machine Learning, vol. 39, pp. 103-134 (2000).

[3] Lee, N.,Caban J., Ebadollahi S., Laine A, “Interactive Segmentation in Multi-Modal Medical Imagery using a Bayesn Transductive Learning Approach”, SPIE Medical Imaging Conference, (2009).

[4] Ebadollahi S., Cooper J., Kaufman D., Levas A., Laine A., DeLaPaz R., Neti C., "Concept-Oriented Access to Longitudinal Multimedia Medical Records: A Case Study in Brain Tumor Management", Workshop on Cross-Media Information Analysis, Extraction and Management, 3rd International Conference on Semantics and Digital Media Technologies, pp. 51-58 (2008).

[5] Lee N., Smith R.T., Laine A., "Interactive Segmentation for Geographic Atrophy in Retinal Fundus Images", IEEE Signal Processing Society, Asilomar Conference on Signals, Systems and Comptuers, (2008).

[6] Lowd D., Domingos P., "Naïve Bayes Models for Probability Estimation", ICML, pp, 529-536 (2005).

[7] McCallum A. and Nigam K., "A Comparison of Event Models for Naive Bayes Text Classification", In AAAI/ICML-98 Workshop on Learning for Text Categorization, Technical Report WS-98-05, AAAI Press, pp. 41-48 (1998).

[8] Jordan M. I., [Learning in Graphical Models], MIT Press, (1998).

[9] Ng A., Jordan M., “On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naïve Bayes”, In NIPS 14 (2002).

[10] Dempster A., Rubin N., "Maximum likelihood from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, vol. 39, no. 1, pp. 1-38 (1977).