The Segmental Boosting Algorithm for Time-series Feature Selection

Pei Yin, Irfan Essa, James M. Rehg
Georgia Institute of Technology
GVU Center / College of Computing
Atlanta, GA 30332-0280 USA
{pyin, irfan, rehg}@cc.gatech.edu

Discriminative feature selection paradigms, *e.g.,* [8, 9] usually consider observation frames in an isolated manner, neglecting temporal dependency in time series. Such temporal relationships provide important information for recognition. We propose Segmental Boosting Algorithm (SBA), which applies feature selection only to the "static segments" of the time-series. It smoothly fills in the gap between the dynamic nature of the time-series data and the static nature of the feature selection methods.

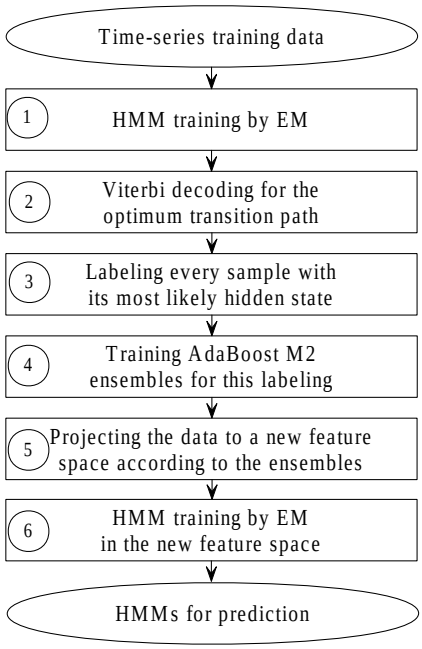

Figure 1. Flow chart of the algorithm.

Hidden Markov model (HMM) has been very successful in interpreting time-series data. HMM builds a causal model for observation sequence $\mathbf{O} = (o_1 o_2 \cdots o_T)$ by introducing corresponding "hidden states" $\mathbf{q} = (q_1 q_2 \cdots q_T)$.

Let $P(q_1) = P(q_1|q_0)$. We denote the parameter of HMM as $\lambda = (\mathbf{a}, \mathbf{b})$ where $\mathbf{a}$ is the parameter for the transition model $P(q_t|q_{t-1})$ and $\mathbf{b}$ is the parameter for the observation model $P(o_t|q_t)$. Assume there are C types of event, the classification is done by selecting the event type with the highest likelihood $c^* = \text{argmax}_{1\le c\le C} P(\mathbf{O}|\lambda_c)$, and $\Lambda = \{\lambda_1, \lambda_2, \cdots, \lambda_C\}$

Define the model distance (dissimilarity) as $D(\lambda_c, \Lambda) = \frac{1}{T_c}[\log P(\mathbf{O}^c|\lambda_c) - \frac{1}{C-1}\sum_{v\neq c} \log P(\mathbf{O}^c|\lambda_v)]$ [5]. Feature selection for time-series data is to choose a subset of features that maximize $D(\lambda_c, \Lambda)$. Assume uninformative prior, it is equivalent to maximize the "margin" of the time-series data $M(\lambda_c, \Lambda) = \frac{1}{T_c}[\log P(\lambda_c|\mathbf{O}^c) - \frac{1}{C-1}\sum_{v\neq c} \log P(\lambda_v|\mathbf{O}^c)]$. Discriminative classifiers with logistic output, *e.g.,* boosting $f(x) = \sum_i h_i(x) = \log P(y = y^*|x) - \log P(y \neq y^*|x)$, is capable of maximizing such margin. However, simply making $(x = \mathbf{O}, y = c)$ is intractable since the length of the observation $\mathbf{O}$ varies from time to time. Alternatively, it is convenient to let $(x = o_t, y = c)$ by assuming temporal independency [8, 9] or fixed dependency [6]. Nevertheless, the neglected temporal dependency in the time-series data conveys important information for recognition. In contrast, we let $(x = o_t, y = q_t)$ for SBA, which respects the temporal dependency while stays tractable. It *decouples* the temporal dependencies from the discriminative feature selection, instead of *discarding* them. The decoupling is derived from Markov property without additional independence assumptions.

Following the Markov property, the likelihood can be decomposed as

$$P(\mathbf{O}|\lambda_c) = \sum_{\mathbf{q}} P(\mathbf{O}|\mathbf{q}, \lambda_c) P(\mathbf{q}|\lambda_c) = \sum_{\mathbf{q}} \prod_{t=1}^{T} P(o_t|q_t, \mathbf{b}(q_t)) P(q_t|q_{t-1}, \mathbf{a}(q_t, q_{t-1}))$$

Thus $M(\lambda_c, \Lambda)$ can be increased with discriminative $P(o_t|q_t, \mathbf{b}(q_t))$. The intuition is to perform feature selection only in the relatively "static segments". The static segments are connected by the temporal transition $P(q_t|q_{t-1}, \mathbf{a}(q_t, q_{t-1}))$. Note that the concept of "hidden state" is still necessary to smooth out the results of the observation model.

Such decoupling has been used in segmental k-means algorithm [2] to obtain good parameter initialization for speech processing. After a random initialization, the set of training sequences are segmented into the optimum state sequence

Table 1. Test Error(%) on Georgia Tech Speech Reading Mocap dataset.

	AdaBoost Only	HMM Only	Boosted HMM [9]	**Segm.Boosting**
Lip Reading	60.18±0.00%	50.36±1.16%	42.56±1.11%	**34.16±1.85%**
	AdaBoost Only	HMM Only	Boosted HMM [9]	**Segm.Boosting**
Speech Rec.	39.69±0.00%	32.30±2.06%	26.54±0.83%	**19.65±1.00%**

Table 2. Test Error(%) on Georgia Tech Speed-Control Gait dataset. First 5 columns are directly from [3]

1-NN DTW	ML(HMM Only)	BML [1]	MixCML [3]	BoostML [4]	BoostedHMM [9]	**Segm.Boosting**
8.38±3.68	11.50±4.78	10.13±3.61	4.00±3.48	11.87±5.11	5.93±6.64	**3.44±1.43**

$\mathbf{q}^* = \text{argmax}_{\mathbf{q}} P(\mathbf{q}|\mathbf{O}, \lambda_c)$ via Viterbi algorithm. Then k-means algorithm is applied to segments $S_q = \{o_t | q_t = q\}, q = 1, 2, \cdots, n$. The observation vectors for each state are clustered into M clusters, where each cluster represents one of the M mixtures of the $\mathbf{b}(q)$ density. $\overline{\mathbf{b}(q)} = \text{argmax}_{\mathbf{b}} \prod_{t, q_t^* = q} P(o_t | q, \mathbf{b}(q))$. This procedure is iterated until convergence.

Similarly, in SBA (Figure 1) we first acquire the optimum state transition path by Viterbi decoding on the HMMs trained with the original features. The HMMs in this step only serve as an estimation of the temporal relationship. To this end, every observation in the training sequence is associated with one hidden state. Then we run AdaBoost on this labeling, and compute a set of ensembles corresponding to every hidden state. Next, a new set of HMMs are trained and tested in the new feature space, outputting the event type with the highest likelihood.

We choose AdaBoost for feature selection because its margin property guarantees that the trajectories of the HMMs become more compact and distinct from each other. Statistical tests show that SBA aggregates the temporal observations in the new feature space with remarkably lower Kurtosis and higher generalized Rayleigh quotient than they are in the original feature space. It indicates that different types of trajectory/motion are more distinguishable.

Experiments on lip reading, gait recognition and speech recognition receive improved accuracy with features selected by SBA. Table 1 shows the classification results on Georgia Tech Speech Reading database [9]. The two classification tasks are to determine the correct phoneme from the lip movement (recorded by motion capture device at 120Hz) or from the speech sound track (recorded at 16KHz, then downsampled to 120Hz) respectively. The database contains over 200,000 samples in 39 phoneme classes. Table 2 shows the classification results on Georgia Tech Speed-Control Gait database [7], to provide a comparison to other concurrent discriminative learning methods for Dynamic Bayesian Networks on a public data source. The training and testing data are processed according to [3] for a fair comparison. The dataset used contains over 90,000 samples in 5 classes (human subjects). Experiments illustrate that SBA achieves lower error by performing feature selection only in the static segments of the time-series data.

We are currently investigating (1) direct estimation of the probability of the observation model for a smoother integration, and (2) whether convergence is required for performance improvement.

References

[1] Y. Jing, V. Pavlovic, and J. Rehg. Efficient discriminative learning of bayesian network classifier via boosted augmented naive bayes. In *International Conference on Machine Learning*, 2005. 2

[2] B. Juang and L. Rabiner. The segmental k-means algorithm for estimating parameters of hidden markov models. pages 1639–1641, 1990. 1

[3] M. Kim and V. Pavlovic. Discriminative learning of mixture of bayesian network classifiers for sequence classification. In *CVPR 06*, pages 268–275. IEEE Computer Society, 2006. 2

[4] V. Pavlovic. Model-based motion clustering using boosted mixture modeling. In *Proc.of IEEE CVPR*, 2004. 2

[5] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ, Printice Hall, 1993. 1

[6] P. Smith and M. Shah N. Lobo. Temporalboost for event recognition. In *Proc. of IEEE ICCV*, pages 733–740, 2005. 1

[7] R. Tanawongsuwan and A. Bobick. Performance analysis of time-distance gait parameters under different speeds. In *4th Internaltional Conference on Audio and Video Based Biometric Person Authentication*, 2003. 2

[8] P. Viola, M. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. In *Proc.of IEEE ICCV*, pages II: 734–741, 2003. 1

[9] P. Yin, I. Essa, and J. Rehg. Asymmetrically boosted HMM for speech reading. In *Proc.of IEEE CVPR*, pages II755–761, 2004. 1, 2

Topics: Estimation, Prediction, and Sequence modeling
Speech and Auditory processing
Preference: Oral/Poster