Context-sensitive Bayesian Classifiers and Application to Mouse Pressure Pattern Classification

Yuan Qi and Rosalind W. Picard
Media Lab, MIT, Cambridge, MA, 02139 , USA
{yuanqi, picard}@media.mit.edu

ABSTRACT

We propose a new context-sensitive Bayesian learning algorithm. This algorithm can utilize different classifier complexities for different contexts/locations and, at the same time, keep the optimality of Bayesian solutions. We apply this algorithm to detecting computer-user mouse pressure patterns, possibly related to frustration, and obtain on average 89.4% test accuracy.

1 Introduction

Motivation Context: e.g., different people, different accents, or different computer-users' habits.

Approach Context-sensitive Bayesian classifiers: softly switching Bayesian classifiers corresponding to different contexts.

- Modeling context: *mixture of Gaussians*
- Training component Bayesian classifiers: *Expectation propagation* [2].
- Combining the output of different component classifiers based on fitting of the context and classification confidence levels.

Benefits Compared to other context learning algorithms [1], this new algorithm has the following advantages:

- soft switching between different hidden contexts
- allowing different complexities, i.e., different kernels, for different component classifiers
- fast training/re-training while keeping the Bayesian optimality

2 Context-Sensitive Expectation Propagation

Given a training set $D = \{D_1, \ldots, D_L\}$ that has L subsets from L possible different contexts

$$D_i = \{(\mathbf{x}_1^i, y_1^i), \ldots, (\mathbf{x}_{N_i}^i, y_{N_i}^i)\}$$

where $i = 1, \ldots, L$, a test data point $\tilde{\mathbf{x}}$ can be classified as follows:

$$p(\tilde{y}|\tilde{\mathbf{x}}, D) = \sum_i p(\tilde{y}|\tilde{\mathbf{x}}, D_i) P(i|\tilde{\mathbf{x}}, D) \qquad (1)$$

Now the problem is to compute:

- $P(i|\tilde{\mathbf{x}}, D) \propto P(\tilde{\mathbf{x}}|D_i)$: modeling context
- $p(\tilde{y}|\tilde{\mathbf{x}}, D_i)$: classification confidence level

These two quantities are computed in the following section 2.1 and 2.2 respectively.

2.1 Modeling Context

Denote $\mathbf{x}^i = \{\mathbf{x}_1^i, \ldots, \mathbf{x}_{N_i}^i\}$, and model $\mathbf{x}^i$ by a Gaussian mixture model H_i, i.e.,

$$p(\mathbf{x}^i) = \sum_{j=1}^{J} \beta_j \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j).$$

where H_i is parameterized by $\{\beta_j, \boldsymbol{\mu}_j, \Sigma_j\}_{j=1}^J$.

- Solve the MAP estimate of $\{\beta_j, \boldsymbol{\mu}_j, \Sigma_j\}_{j=1}^J$ by the Expectation Maximization algorithm.
- Approximate $p(\tilde{\mathbf{x}}|D_i)$:

  $$p(\tilde{\mathbf{x}}|D_i) = \int_{H_i} p(\tilde{\mathbf{x}}|H_i) p(H_i|D_i) \approx p(\tilde{\mathbf{x}}|\hat{H}_i)$$

  where $\hat{H}_i$ corresponds to the EM estimate.

$$P(i|\tilde{\mathbf{x}}, D)) \approx \frac{p(\tilde{\mathbf{x}}|\hat{H}_i)}{\sum_i p(\tilde{\mathbf{x}}|\hat{H}_i)} \qquad (2)$$

2.2 Training Component Classifiers

Classifier $\mathbf{w}$: classify a point $\mathbf{x}$ according to $y = \mathrm{sign}(\mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}))$

Function $\phi(\mathbf{x}_i)$: Gaussian cdf introduces nonlinearity into the classification boundary

Likelihood: $p(\mathbf{y}|\mathbf{w}, X) = \prod_i p(y_i|\mathbf{w}, \mathbf{x}_i) = \prod_i \Psi(y_i \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_i))$ where $\Psi(a) = 1$ if $a > 0$ and $\Psi(a) = -1$ if $a \leq 0$.

The predictive distribution of $\mathbf{x}_{\mathbf{N+1}}$:

$$p(y_{N+1}|\mathbf{x}_{N+1}, \mathbf{y}) = \int_{\mathbf{w}} p(y_{N+1}|\mathbf{w}, \mathbf{x}_{N+1}) p(\mathbf{w}|\mathbf{y})$$
$$\approx P(y_{N+1}|\mathbf{x}_{N+1}, \langle \mathbf{w} \rangle) \qquad (3)$$

where $\langle \mathbf{w} \rangle$ denotes the posterior mean of $\mathbf{w}$, called the Bayes Point, efficiently computed by Expectation Propagation [2].

Computing $p(\tilde{y}|\tilde{\mathbf{x}}, D_i)$ as follows:

$$p(\tilde{y}|\tilde{\mathbf{x}}, D_i) = \int_{\mathbf{w}} p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{w}) p(\mathbf{w}|D_i) = \Psi(\tilde{z}) \qquad (4)$$

$$\text{where } \tilde{z} = \frac{(\tilde{y}\mathbf{m}_w)^T \boldsymbol{\phi}(\tilde{\mathbf{x}})}{\sqrt{\boldsymbol{\phi}(\tilde{\mathbf{x}}) \mathbf{V}_w \boldsymbol{\phi}(\tilde{\mathbf{x}})}} \qquad (5)$$

where $\mathbf{m}_w$ and $\mathbf{V}_w$ are the posterior mean and variance of the classifier $\mathbf{w}$.

2.3 Combining Component Classifiers

Compute $p(\tilde{y}|\tilde{\mathbf{x}}, D)$: eq. (1), (4), and (2).

Classify $\tilde{y}$: if $p(\tilde{y} = 1|\tilde{\mathbf{x}}, D) > p(\tilde{y} = -1|\tilde{\mathbf{x}}, D)$, then $\tilde{y} = 1$; otherwise, $\tilde{y} = -1$.

3 CS-EP v.s. ME [4]

Compare CS-EP to Mixture of Experts (ME):

- More general Gating function than ME.
- Training each component classifier once while ME requires iterative training.
- Kernelized Bayes Classifier instead of logistic regression in ME.

4 Mouse Pressure Pattern Classification

Classifing mouse pressure signals from different subjects (S1 to S5) into two categories:

- pressure patterns that arise when all is going well,
- those that arise following potentially frustrating events, like usability bugs.

Hidden Context: user identity
Experimental results:

Kernel Width	CS-EP Test Error (%)						SVM Err. (%)
	S1	S2	S3	S4	S5	Avg.	
0.1	7.1	8.0	12.3	11.8	28.7	13.6	14.3
0.5	8.0	10.0	9.0	5.9	23.0	11.2	11.4
0.1 & 0.5	7.1	8.0	9.0	5.9	23.0	10.6	N/A

CS-EP uses less computation, memory, and time, and achieves lower error rates than SVM on the average. Combining different kernel widths improves CS-EP's test performance.

5 Future Work

- Hierarchical Modeling
- Handling unbalanced classes

References

[1] M. Harries and K. Horn. Learning stable concepts in domains with hidden changes in context. In *13th ICML, workshop on Learning in Context Sensitive Domains*, 1996.

[2] T. P. Minka. Expectation propagation for approximate Bayesian inference. In *UAI'01*, 2001.

[3] Y. Qi, C. Reynolds, and R. W. Picard. The Bayes point machine for computer-user frustration detection via pressuremouse. In *PUI 2001*.

[4] L. Xu, M. I. Jordan, and G. E. Hinton. An alternative model for mixtures of experts. *NIPS 7*, pages 633–640. 1995.