ADAPTIVE EEG SIGNAL CLASSIFICATION USING STOCHASTIC APPROXIMATION METHODS

Shiliang Sun, Man Lan, Yue Lu

Department of Computer Science and Technology, East China Normal University
500 Dongchuan Road, Shanghai 200241, P. R. China
slsun@cs.ecnu.edu.cn, lanman.sg@gmail.com, ylu@cs.ecnu.edu.cn

ABSTRACT

Classification of time-varying electrophysiological signals is an important problem in the development of brain-computer interfaces (BCIs). Designing adaptive classifiers is a potential way to address this task. In this paper, Bayesian classifiers with Gaussian mixture models (GMMs) are adopted as the decision rule to classify electroencephalogram (EEG) signals. The stochastic approximation method (SAM) is used as the specific gradient descent method for updating the parameters of mean values and covariance matrices in the distribution of GMMs, where the parameters are simultaneously updated in a batch mode. Experimental results using data from a BCI show that the stochastic approximation method is effective for EEG classification tasks.

Index Terms— Bayesian classifier, brain-computer interface (BCI), EEG signal classification, Gaussian mixture model (GMM), stochastic approximation method (SAM)

1. INTRODUCTION

During the last few years, the research pace of brain-computer interfaces (BCIs) has quickened greatly. A BCI is a channel for communication and control, which does not depend on the brain's traditional output pathways of peripheral nerves and muscles [1, 2]. Its potential applications include restoring functions to those with motorial disabilities, alarming paroxysmal diseases, manipulating human's control in inhospitable even dangerous environments, etc [2]. Due to its intrinsic complexity, research on a BCI is an interdisciplinary field with neuroscience, psychology, engineering, clinical rehabilitation, and computer science included.

Electrophysiological signal classification which translates people's intents to external device commands is a central but challenging task in a BCI. It recently attracts many attentions from the signal processing and machine learning community [2, 3, 4, 5]. The focus of this paper is on the problem of electrophysiological signal classification arising in the current BCI research. The signals considered here are electroencephalograms (EEGs), which are electrophysiological signals recorded in terms of the electroencephalography with electrodes non-invasively placed on the human scalp. They reflect electrical brain activities essentially generated by the underlying neurons in the cortex.

For a BCI in use, adaptive classification mechanisms are necessary because EEG signals are typically instable, namely they change over time due both to biological and to technical factors. Biologically, they change due to user fatigue and attention, due to disease progression, and with the process of training. Technically, they change due to amplifier noises, ambient noises, and the variation of electrode impedances [2]. It will be challenging and often impossible even for the same user to adopt a classifier trained on the first day to classify data recorded during following days without retraining. Millán has shown that two different mental tasks, imagination of left and right hand movements respectively, can have closer power maps than the same task during two consecutive recording sessions [6]. The spontaneous variability of EEG recordings between experimental sessions makes it difficult to classify different EEG signals accurately, and necessitates on-line learning to improve the performance of the classifiers.

Although some methods including linear and nonlinear have been proposed for EEG signal classification, e.g., Fisher discriminant analysis, support vector machines, artificial neural networks, hidden Markov models [3, 4, 7, 8, 9, 10], they are stationary in nature and can not effectively tackle the problem of classifying time-varying EEG signals. Up to the present there is little work in the current literature addressing the problem of on-line EEG signal classification. The articles [6, 11, 12] are among the earliest ones discussing the problem of online EEG signal classification based on Bayesian classifiers and stochastic gradient methods (SGMs). In this paper, we propose to use the stochastic approximation method (SAM) for learning adaptive Bayesian classifiers. Experimental results on EEG signal classification and comparisons with the basic SGM are implemented to evaluate the feasibility of the proposed method.

Thanks to the National Natural Science Foundation of China and Shanghai Educational Development Foundation for funding respectively under Project 60703005 and Project 2007CG30.

2. ADAPTIVE BAYESIAN CLASSIFIERS WITH GMMS

As the articles [6, 11, 12] suggest, we adopt Bayesian classifiers to carry out on-line EEG signal classification. Bayesian classifiers assign the label of a sample to the class which has the largest posterior probability. For the on-line applications of BCIs, the recorded data are increasing dynamically. The adaptive Bayesian classifier considers how to adaptively update parameters using the new added samples, and then classify the forthcoming samples.

Suppose there are N training samples which come from K classes, and each class denoted by C_k has the prior probability $P(C_k)$ $(k = 1, ..., K)$, s.t., $\sum_{k=1}^{K} P(C_k) = 1$. Under the framework of finite GMMs, the conditional distribution of each class can be approximated by the weighted combination of several Gaussian distributions [13], i.e.,

$$p(x|C_k) \cong \sum_{i=1}^{N_k} a_k^i G(x|\mu_k^i, \Sigma_k^i), s.t., \Sigma_{i=1}^{N_k} a_k^i = 1,\ a_k^i > 0\ , \tag{1}$$

where $G(x|\mu_k^i, \Sigma_k^i)$ is the Gaussian distribution with the mean value μ_k^i and the covariance matrix Σ_k^i, and N_k is the number of Gaussian distributions enclosed in a GMM. a_k^is are combinational weights of the corresponding Gaussian distributions.

According to Bayes formula [14], given a sample x the posterior probability of the class C_k can be transformed as

$$P(C_k|x) = \frac{P(C_k)\sum_{i=1}^{N_k} a_k^i G(x|\mu_k^i, \Sigma_k^i)}{\sum_{j=1}^{K} P(C_j)\sum_{i=1}^{N_j} a_j^i G(x|\mu_j^i, \Sigma_j^i)}\ . \tag{2}$$

Let us denote the N available samples as $\{x_n, y_n\}$ $(n = 1, ..., N)$, where x_n is the feature vector of the n^{th} sample, and y_n is the corresponding label vector having K possible states. If $x_n \in C_k$, y_n has the form of e_K^k (using the 1-of-K coding mechanism), that is,

$$y_n \triangleq e_K^k = [0, \ldots, 1^{(k)}, \ldots, 0]_{(K)}^{\top}. \tag{3}$$

Similarly, denote $\hat{y}_n$ as the outcome of the Bayesian classifier for the input x_n,

$$\hat{y}_n \triangleq [P(C_1|x_n), P(C_2|x_n), \ldots, P(C_K|x_n)]^{\top}\ . \tag{4}$$

Under the criterion of least mean square error, the optimization objective function for estimating parameters of the Bayesian classifiers is

$$\min J(\Theta) \triangleq \min E\{\|y_n - \hat{y}_n\|^2\}\ , \tag{5}$$

where the variable Θ represents any of the parameters N_k, a_k^i, μ_k^i, Σ_k^i given in (1). Since the current objective function involves the computation of mathematical expectation, the corresponding optimization task is called the stochastic optimization problem [16].

For the convenience of later analysis, parameters N_k, a_k^i are presumed to be given or obtained from training data, while parameters μ_k^i, Σ_k^i have the most general forms (μ_k^i is a general column vector, and Σ_k^i is a qualified covariance matrix with symmetric and positive definite properties) whose values will be obtained through on-line update.

3. METHODS FOR GRADIENT DESCENT

For the adaptive parameter update in the Bayesian classifiers, gradient descent methods, which are also called the steepest descent methods, can be used. However, as the mathematical expectation in (5) can not be obtained precisely in real applications, it is infeasible to directly calculate the gradients of related parameters for the stochastic optimization problem. Some approximation should be applied. The SGM and the SAM are two feasible variations of gradient descent methods for approximating the stochastic optimization problem.

3.1. SGM

The SGM alters the objective function in (5) by replacing the computation of mathematical expectation with the instantaneous value $\|y_n - \hat{y}_n\|^2$. The gradient $\nabla_\Theta \|y_n - \hat{y}_n\|^2$ of the new objective function is called the instantaneous gradient. Now the gradient descent algorithm can be written as:

$$\Theta(n) = \Theta(n-1) - \mu(n) \cdot \nabla_\Theta \|y_n - \hat{y}_n\|^2\ , \tag{6}$$

where Θ represents the parameters to be updated, and $\mu(n)$ is the learning rate at the time instant n [15]. For a new sample, the SGM first calculates the instantaneous gradient, and then carries out adaptive update using (6).

3.2. SAM

As an alternative for (5), we can also adopt a considerable number of samples to approximate the calculation of mathematical expectation. The optimization problem becomes

$$\min J_N(\Theta) \triangleq \min \frac{1}{N}\sum_{n=1}^{N} \|y_n - \hat{y}_n\|^2\ , \tag{7}$$

which is called the deterministic optimization problem as it doesn't include the mathematical expectation [16, 17]. The gradient descent method solving the problem is called the SAM, which is given by

$$\Theta(n) = \Theta(n-1) - \mu(n) \cdot \frac{1}{N}\sum_{i=1}^{N} \nabla_\Theta \|y_i - \hat{y}_i\|^2\ , \tag{8}$$

where parameters Θ and $\mu(n)$ have the same meanings as before.

The SAM is a batch processing algorithm, which employs a pool of samples to calculate gradients and update parameters. Therefore, it can in principle discover parameters which

are suitable to the current data pool. With the continually added samples, the SAM can renew the data pool used for gradient computation, and consequently update the corresponding parameters adaptively. Note that in order to facilitate the selection of the learning rates, the gradients $\nabla_\Theta \|y_n - \hat{y}_n\|^2$ in (6) and $\frac{1}{N}\sum_{i=1}^{N} \nabla_\Theta \|y_i - \hat{y}_i\|^2$ in (8) are usually normalized to be identity vectors.

3.3. Qualitative comparison

The SGM and the SAM are applicable to update parameters for learning adaptive Bayesian classifiers. However, the difference on the number of samples used to calculate gradients may induce distinct results of parameter learning. Since the SGM only adopts one sample for subsequent parameter update and the single sample may not satisfactorily represent the distribution of future samples, the learned parameters may not generalize well to unknown samples. Making it worse, the negative influence from possible noises can deteriorate the accuracy of the computed gradients to a large extent.

Different from the SGM, the SAM integrates more than one sample in computing gradients. Thereby the negative influence from noises is largely reduced. Furthermore, if the data pool is representative, the resultant classifier parameters will generalize well to unknown samples. Therefore, here we adopt the SAM for learning adaptive Bayesian classifiers.

The formulation for the instantaneous gradient $\nabla_\Theta \|y_n - \hat{y}_n\|^2$ is needed to apply whether the SGM or the SAM. Since it was already derived in [12], here we simply adopt the result to carry out subsequent parameter update.

4. CASE STUDY

4.1. Data description and parameter setup

The data used in this paper are provided by the IDIAP Research Institute of Switzerland as a benchmark for algorithm evaluation. They are EEG recordings taken from normal subjects during three mental imagery tasks. The mental tasks are imagination of repetitive left hand movements (class C_1), imagination of repetitive right hand movements (class C_2) and generation of different words beginning with the same random letter (class C_3). Data from the first two subjects (denoted by S1 and S2 respectively) are used. For a given subject, there are four non-feedback sessions recorded. After spatial filtering and power spectral density estimation, the raw EEG signals are converted to 96-dimensional feature vectors with every 12 entries coming from one of eight centro-parietal electrodes (EEG signals recorded over this region can reflect the discriminative activities of brain's sensorimotor cortices). The numbers of samples in the four sessions for subjects S1 and S2 are respectively 3488-3472-3568-3504, and 3472-3456-3472-3472 [18]. In this paper, the 96 dimensional precomputed features are adopted to simulate on-line classification.

Six data sets are constructed from the above data for algorithm evaluation. Each data set consists of a training set and a test set. The first three are formed using the data of four sessions from S1, and so on, for a total of six data sets. Specifically, data sets 1, 2, 3 are respectively composed of session pairs $1 \sim 2$, $2 \sim 3$, $3 \sim 4$ of S1. Data sets 4 to 6 are constituted with a similar style from subject S2. For each pair of sessions, the former session serves as the training set, and the latter the test set.

Parameters $P(C_k), N_k$ and a_k^i in the adaptive Bayesian classifiers are taken as the same setup in [6], that is, $P(C_k) = \frac{1}{3}$, $N_k = 4$ and $a_k^i = \frac{1}{4}$ $(k = 1, 2, 3; i = 1, 2, 3, 4)$. To reduce the parameters to be estimated, principal component analysis is adopted to reduce the feature dimension by reserving 90% energy on each data set [14]. By running the SAM on a training set, we can obtain the estimated values of μ_k^i and Σ_k^i, which are taken as the initial configurations for the on-line update of classifier parameters on the corresponding test set. The termination condition for learning μ_k^i and Σ_k^i on a training set is either 100 steps of iteration are reached or parameters converge.

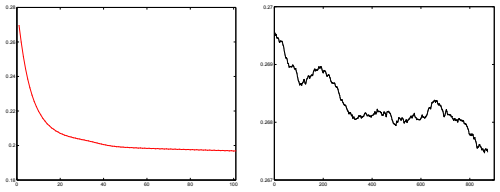
4.2. Quantitative comparison of SAM and SGM

To quantitatively evaluate the performance of the SAM and the SGM for gradient descent, we now compare these two methods on the available training sets. Without loss of generality, the size of each training set is reduced to one-fourth of the original size by down-sample processing.

The objective function of the SAM is taken as the mean square error of all samples in the training set. The learning rates for updating μ_k^i and $(\Sigma_k^i)^{-1}$ are fixed as the same value, which is selected within the small range of $\{1e-1, 1e-2, 1e-3, 1e-4\}$. The range is chosen empirically based on previous knowledge on adaptive learning. Experimental results indicate that if the learning rate is $1e-1$ or $1e-2$, the objective function will not converge, while if the learning rate is $1e-4$, it will converge extremely slowly. Therefore, the learning rate is finally set as $1e-3$. The maximal steps of iteration are fixed as 100.

The objective function of the SGM is the same with that of the SAM. During the phase of parameter update, each sample in the training set serves as the input in turn, and is used only once. The learning rate is selected from a small range $\{1e-3, 1e-4, 1e-5, 1e-6\}$. Due to the same reason for selecting learning rates in the SAM, the learning rate in the SGM is finally set as $1e-5$.

Fig. 1 shows the curves of the values of objective functions with increasing iteration steps by the two gradient descent methods on data sets 1, and 4. The curves on other data sets are quite similar. The result reveals that given finite training samples the SAM can always decrease the values of objective functions, while the SGM can not do this and the corresponding curves don't have a steady tendency. Thereby,



(a) Data set 1 (b) Data set 1

(c) Data set 4 (d) Data set 4

Fig. 1. The values of objective functions with increasing iteration steps. The left two figures are obtained by the SAM, while the right two figures are obtained by the SGM.

Table 1. The classification accuracies (%) obtained by the adaptive (Ada) and off-line (Off) Bayesian classifiers

	Data set					
	1	2	3	4	5	6
Off	73.13	65.17	73.48	45.77	48.65	52.13
Ada	**73.20**	**65.53**	**73.80**	**46.03**	**49.16**	**52.45**

the effectiveness of the SAM is verified. This is the motive of using the SAM in learning adaptive Bayesian classifiers.

4.3. Classification performance of the SAM

The data of the first two minutes from each test data set are adopted to update the parameters of the Bayesian classifiers. The rest of the data are used to test classification performance of the learned classifiers. The learning rate is set as $1e-3$, which is selected above on the training set. The iteration step is simply taken to be 1.

The classification accuracies respectively obtained by the adaptive Bayesian classifiers and the corresponding classifiers without update (off-line) on the test data sets are given in Table 1. We can find that, on all the data sets the accuracies of the adaptive classifiers are better than those of the off-line classifiers. Although the current performance improvement is small, it can be expected to enlarge a lot if multiple iteration steps rather than only 1 as used here are allowed.

5. CONCLUSIONS

In this paper, we address the problem of on-line classifying EEG signals and propose to use the SAM for updating parameters of the Bayesian classifiers. Experiments show the effectiveness and potential of the SAM. In the future, investigating the performance of multiple iterations for the SAM using more data sets and exploring the feasibility of other gradient descent methods would be interesting.

6. REFERENCES

[1] M.A.L. Nicolelis, "Actions from thoughts," *Nature*, vol. 409, no. 6818, pp. 403–407, 2001.

[2] J.R. Wolpaw, N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[3] B. Blankertz, G. Curio, and K.R. Müller, "Classifying single trial EEG: Towards brain computer interfacing," in *Advances in Neural Information Processing Systems*, vol. 14, pp. 157–164, 2002.

[4] T.N. Lal, T. Hinterberger, G. Widman, M. Schröder, J. Hill, W. Rosenstiel, C. E. Elger, B. Schölkpf, and N. Birbaumer, "Methods towards invasive human brain computer interfaces," in *Advances in Neural Information Processing Systems*, vol. 17, pp. 737–744, 2005.

[5] S. Sun and C. Zhang, "An optimal kernel feature extractor and its application to EEG signal classification," *Neurocomputing*, vol. 69, no. 13-15, pp. 1743–1748, 2006.

[6] J.R. Millán, "On the need for on-line learning in brain-computer interfaces," in *Proc. Int. Joint Conf. Neural Networks*, pp. 2877–2882, 2004.

[7] C.W. Anderson, E.A. Stolz, and S. Shamsunder, "Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 277–286, 1998.

[8] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, "BCI competition 2003-data set IIb: Support vector machines for the P300 speller paradigm," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1073–1076, 2004.

[9] Y. Wang, Z. Zhang, Y. Li, X. Gao, S. Gao, and F. Yang, "BCI competition 2003-data set IV: An algorithm based on CSSD and FDA for classifying single-trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1081–1086, 2004.

[10] S. Zhong, and J. Ghosh, "HMMs and coupled HMMs for multi-channel EEG classification," in *Proc. Int. Joint Conf. Neural Networks*, pp. 1154–1159, 2002.

[11] J.R. Millán, F. Renkens, J. Mouriño, and W. Gerstner, "Noninvasive brain-actuated control of a mobile robot by human EEG," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1026–1033, 2004.

[12] S. Sun, and C. Zhang, "Learning on-line classification via decorrelated LMS algorithm: application to brain-computer interfaces," *Lecture Notes in Computer Science*, vol. 3735, pp. 215–226, 2005.

[13] G. Mclachlan, and D. Peel, *Finite Mixture Models*. New York: Wiley, 2000.

[14] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, 2nd ed. New York: John Wiley & Sons, 2000.

[15] G.O. Glentis, K. Berberidis, and S. Theodoridis, "Efficient least squares adaptive algorithms for FIR transversal filtering," *IEEE Signal Process. Mag.*, vol. 16, no. 4, pp. 13–41, 1999.

[16] H. Kusner, and G. Yin, *Stochastic Approximation Algorithms and Applications*. New York: Springer-Verlag, 1997.

[17] X. Zhang, *Matrix Analysis and Applications*. Beijing: Tsinghua University Press, 2004.

[18] S. Chiappa, and J.R. Millán, Data set V <mental imagery, multi-class>. Available: http://ida.first.fraunhofer.de/projects/bci/competition_iii/desc_V.html