Robust Learning of Mixture Models and Its Application on Trial Pruning for EEG Signal Analysis

Boyu Wang, Feng Wan, Peng Un Mak, Pui In Mak, and Mang I Vai

Department of Electrical and Electronics Engineering,
Faculty of Science and Technology,
University of Macau

Abstract. This paper presents a novel method based on deterministic annealing to circumvent the problem of the sensitivity to atypical observations associated with the maximum likelihood (ML) estimator via conventional EM algorithm for mixture models. In order to learn the mixture models in a robust way, the parameters of mixture model are estimated by trimmed likelihood estimator (TLE), and the learning process is controlled by temperature based on the principle of maximum entropy. Moreover, we apply the proposed method to the single-trial electroencephalography (EEG) classification task. The motivation of this work is to eliminate the negative effects of artifacts in EEG data, which usually exist in real-life environments, and the experimental results demonstrate that the proposed method can successfully detect the outliers and therefore achieve more reliable result.

Keywords: Deterministic annealing, mixture models, robust learning, trial pruning, EEG signals.

1 Introduction

A brain computer interface (BCI) is a system forming a direct connection between brain and machine, which enables individuals with severe motor disabilities to have effective control over external devices without using the traditional pathways as peripheral muscle or nerves [1-3]. The brain activities are often recorded noninvasively by electroencephalogram (EEG), which has excellent temporal resolution and usability, and the EEG signal is therefore a popular choice for BCI research. In order to control an EEG-based BCI, the user must produce different brain activity patterns, which are recorded by electrodes on the scalp, and then features are extracted from the EEG signals and translated into the control commands. In most existing BCIs, this translation relies on a classification algorithm [4], [5].

Finite mixture models, in particular Gaussian mixture models (GMMs) [6] have been applied to EEG signal analysis in BCI system due to their computational tractability, ease to implement, and capability of representing arbitrarily complex probability density function with high accuracy. In [7] and [8], the mixture of Gaussian was introduced as the online classifier and the parameters were updated in a simulated online scenario. In [9] a GMM-based classifier was used to separate the signal into different classes of mental task, where adaptation is concerned by using a

L. Cao et al. (Eds.): PAKDD 2011 Workshops, LNAI 7104, pp. 408–419, 2012.

supervised method. Similarly, [10] and [11] proposed an online GMM classifier via the decorrelated least mean square (DLMS) algorithm. On the other hand, GMMs can be also applied to model the features extracted from EEG data in which the rest or active state of brain signals are modeled so that the changes in EEG signal can be detected rather than classified [12], [13].

The conventional approach to learning the parameters of mixture models is maximum likelihood (ML) estimator via EM algorithm [14]. However, a well-known problem of the ML estimator via conventional EM algorithm for GMMs is its sensitivity to atypical observations. On the other hand, noise is ubiquitous in EEG signals due to the factors such as measurement inaccuracies, physiological variations in background EEG, muscle and eyes blink artifacts. Therefore, contaminated samples in EEG data should be pruned to achieve a reliable classification result. Unfortunately, none of the GMM-based EEG analysis algorithm, to our best knowledge, considered the negative effects of the outliers in EEG data.

In machine learning community, one approach to detect the outliers is to fit the model based on maximum trimmed likelihood (MTL) to select a subset of the data, on which the mixture models are trained to the majority of the data, whereas the remaining data which do not follow the models are viewed as anomalous data. The ML estimator can be viewed as a special case of MTL estimator. The resulting estimation of the parameters obtained by trimmed likelihood estimator (TLE) is usually more robust, and therefore can be used for outlier detection [17] [18]. One drawback of this approach, however, is that it is a local algorithm, and therefore usually gets trapped in local optima with a poor estimation.

The motivation of this paper is to go a step further along in this research direction, that is, to develop a robust learning algorithm for mixture models. In particular, we propose a deterministic annealing (DA) learning approach for robust fitting of GMMs. The GMMs are learned based on MTL via EM algorithm and the learning process is controlled by annealing temperatures, leading gradual optimization of the objective function, so that the local optima problem of MTL can be avoided. As a result, the outliers can be automatically detected so that the estimation of parameters of GMMs is more robust and reliable.

The reminder of this paper is organized as follows. The robust learning algorithm is developed in Section 2, and experiments on both synthetic and benchmark real data sets are reported in Section 3. In Section 4, we apply our method on EEG signal analysis. The conclusions are provided in Section 5.

2 Deterministic Annealing for Robust Learning

2.1 Mixture Models, Trimmed Likelihood Estimator and FAST-TLE

Given a data set $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, consisting of N independent identical distributed (i.i.d.) observations of a random d-dimensional variable $\mathbf{x}$. If it follows a K-component finite mixture distribution, its probability density function (pdf) can be given by:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, p\left(\mathbf{x} \mid \boldsymbol{\theta}_k\right), \text{ with } 0 \le \pi_k \le 1 \text{ and } \sum_{k=1}^{K} \pi_k = 1 \tag{1}$$

where π_k is the mixing coefficient, and $\boldsymbol{\theta}_k$ is the parameter set for the kth component.

Define $\boldsymbol{\Theta} \equiv \{\pi_1, ..., \pi_K, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K\}$ as the complete set of the parameters specifying the mixture model. The ML estimate of the optimal set of the parameters is defined as a maximum of the log-likelihood function:

$$\log p(\mathbf{X} \mid \boldsymbol{\Theta}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \Theta) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \, p\left(\mathbf{x}_n \mid \boldsymbol{\theta}_k\right) \tag{2}$$

It is well known that the ML estimator cannot be obtained in a closed form. Hence, we need to resort the optimization techniques, and one common choice is the EM algorithm, which is an iterative procedure to find the ML estimator of the parameter set of a probability. For more detailed description of the EM algorithm see [6], [14].

To estimate GMMs in a robust way, one approach is to calculate the MTL solution, which is given by

$$\log p_{TL}(\mathbf{X} \mid \boldsymbol{\Theta}) = \sum_{n=1}^{N} \omega_n \log p(\mathbf{x}_n \mid \boldsymbol{\Theta}) = \sum_{n=1}^{N} \omega_n \log \sum_{k=1}^{K} \pi_k \, p\left(\mathbf{x}_n \mid \boldsymbol{\theta}_k\right) \tag{3}$$

where $\omega_n \in \{0,1\} \ \forall \ n = 1, ..., N$, and $\sum_{n=1}^{N} \omega_n = M$ is the indicator describing which of the N observations is viewed as a typical sample, so that $\mathbf{x}_n$ is detected as an outlier when $\omega_n = 0$, and $\omega_n = 1$ for a typical sample, which contributes to the log-likelihood function. M is the trimming parameter indicating the removal of (N-M) samples which cannot be fitted well by any component of the mixtures, and MTL degenerates to ML when N=M.

The FAST-TLE algorithm was proposed in [19] to get an approximation of TLE solution for generalized linear models (GLMs), and was extended for robust fitting of mixture model in [18]. The FAST-TLE algorithm consists of two steps called the trial step and a refinement step. However, this algorithm is a local method, and therefore may get trapped in local optima resulting in a poor estimation.

2.2 Deterministic Annealing Outlier Detection

To avoid the local optima problem with FAST-TLE, we resort the deterministic annealing (DA) approach in which the optimization problem is reformulated as that of seeking the probability distribution that minimizes the application-specific cost subject to a constraint of randomness of the solution. During the annealing process, the algorithm tracks the minimum while the temperature is gradually lowered so that many shallow local optima can be avoid, and finally achieves the hard (nonrandom) solution as the temperature approaches to zero [20], [21].

In the light of TLE and DA algorithm, we consider the following objective function:

$$F(\boldsymbol{\Theta},\omega) = -\sum_{n=1}^{N} \omega_n \log p(\mathbf{x}_n \mid \boldsymbol{\Theta}) - TH_\omega \tag{4}$$

under the constraints $\sum_{n=1}^{N} \omega_n = M$, $\sum_{k=1}^{K} \pi_k = 1$, and

$$H_\omega = -\sum_{n=1}^{N} \omega_n \log \omega_n \tag{5}$$

where T is the Lagrange multiplier, which is analogous to the temperature in statistics physics, H_ω is the Shannon entropy, which represents a specified level of randomness. At high value of T, the objective function is very smooth and we mainly maximize the entropy, with $\sum_{n=1}^{N} \omega_n = M$, yielding $\omega_n = M / N$, i.e., each samples is equally treated, and the MTL is therefore equivalent to ML. Hence, we have

$$\min_{\boldsymbol{\Theta},\omega} \lim_{T\to\infty} F(\boldsymbol{\Theta},\omega) = \max_{\boldsymbol{\Theta}} \log p(\mathbf{X} \mid \boldsymbol{\Theta}) = \max_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \boldsymbol{\Theta}) \tag{6}$$

As T is gradually lowered, the influence of log-likelihood function is increasing, which makes the solution of ω_n harder and harder. Finally, as T approaches to zero, the optimization is carried out directly on the trimmed log-likelihood function, forcing ω_n to either zero or one, which yields the MTL

$$\min_{\boldsymbol{\Theta},\omega} \lim_{T\to 0} F(\boldsymbol{\Theta},\omega) = \max_{\boldsymbol{\Theta}} \log p_{TL}(\mathbf{X} \mid \boldsymbol{\Theta})\big|_{\omega_n \in \{0,1\}} = \max_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \omega_n \log p(\mathbf{x}_n \mid \Theta)\big|_{\omega_n \in \{0,1\}} \tag{7}$$

The motivation of the DA based learning procedure is that there is no guarantee that the selected subset in the early stage of learning is near the true one. Therefore, all of the samples should be equally treated at early stage, and the constraint is gradually relaxed during the learning process to increase the effect of the selection of subset, so that the global (at least a better local) optimal solution could be achieved.

For GMMs, to maximize (4), given the fixed $\{\omega_n\}$, we have

$$\pi_k = \frac{1}{M} \sum_{n=1}^{N} \omega_n p(k \mid \mathbf{x}_n), \quad \boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} \omega_n p(k \mid \mathbf{x}_n) \mathbf{x}_n}{\sum_{n=1}^{N} \omega_n p(k \mid \mathbf{x}_n)} \tag{8}$$

$$\boldsymbol{\Sigma}_k = \frac{\sum_{n=1}^{N} \omega_n p(k \mid \mathbf{x}_n)(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T}{\sum_{n=1}^{N} \omega_n p(k \mid \mathbf{x}_n)}$$

where

$$p(k \mid \mathbf{x}_n) = \frac{\pi_k N(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{9}$$

It can be observed from (8) that the larger the value of ω_n is, the more the corresponding sample contributes to the estimation of the parameters. As T approaches to zero, ω_n skews either to one or zero, indicating whether the sample is viewed as typical or eliminated as an outlier.

For the parameters $\{\omega_n\}$, we minimize the following objective function

$$\omega_n = \arg\min_{\omega_n} \left\{ \begin{array}{l} F_{RDAEM}(\boldsymbol{\Theta}, \boldsymbol{\omega}, \upsilon) \\ \text{s.t. } \sum_{n=1}^{N} \omega_n = M,\ \omega_n \in [0,1] \end{array} \right\} \tag{10}$$

This bound-constrained convex optimization can be solved by cvx, a Matlab package for solving convex program [22].

A description of the proposed algorithm for GMMs is summarized in Fig. 1.

Algorithm: *Deterministic Annealing Based Approach for Outlier Detection*

Input: Data Matrix $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, scaling factor α, $T_{\max}$, and $T_{\min}$

Output: Optimal mixture model $\boldsymbol{\Theta}$, and outlier indicators $\{\omega_n\}$

Procedure:

Initialize the parameter set of the mixture models $\boldsymbol{\Theta}$, the indicator $\omega_n = M / N$, and minimum temperature $T_{\min}$,. Set $T = T_{\max}$, $t = 0$.

Repeat

Repeat

$t = t+1$

E-Step:

Calculate $p(k \mid \mathbf{x}_n)$ according (9)

M-Step:

Update $\{\pi_k\}$, $\{\boldsymbol{\mu}_k\}$, and $\{\boldsymbol{\Sigma}_k\}$ according to (8)

Update $\{\omega_n\}$

Until a stop criterion is met.

$T = \alpha T$ $(0 < \alpha < 1)$.

Until $T < T_{\min}$,

Return the model parameter set $\boldsymbol{\Theta}$ and $\{\omega_n\}$

Fig. 1. Deterministic Annealing Based Approach for Outlier Detection

In the proposed algorithm, $T_{\max}$ is set to 100, and $T_{\min}$ is set in the range of [0.005 0.01]. The choice of scaling factor α involves a tradeoff between execution time and the risk of poorer performance. In practical application, $\alpha \in [0.8\ 0.9]$ can achieve satisfactory results.

The stop criterion can be a maximum number of EM cycles or a convergent indicator. In general, if the algorithm is executed until convergence or the maximum number is set to a large value, the computation time is longer, and the algorithm may get trapped in local minima in the early stage of learning. Therefore, the criterion is set to be execution of 10 to 20 EM cycles at each temperature in our experiment.

In summary, the learning process consists of repeated E-step and modified M-step while gradually lowering the temperature, and its monotonicity in objective function is obvious. When the temperature approaches to zero, the method degenerates to FAST-TLE algorithm, of which the monotonicity has been proved in [19]. The convergence property of deterministic annealing has also been discussed in [21], [23].

3 Experiments

3.1 Synthetic Data Sets

The first example is a synthetic dataset which consists of 100 samples from three Gaussian components with equal mixing coefficients, and the parameter set of each component is given by

$$\boldsymbol{\mu}_1 = [0,\ 3]^T,\ \boldsymbol{\mu}_2 = [3,\ 0]^T,\ \boldsymbol{\mu}_3 = [-3,\ 0]^T$$

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},\ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ \boldsymbol{\Sigma}_3 = \begin{bmatrix} 2 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}$$

In addition, 50 noise points generated from a uniform distribution within [-10, 10] on each dimension are added to the typical samples, which is similar to the data set discussed in [1], [15], [18]. Thus, the total number of samples N = 150, and the number of typical samples M = 100. The obtained samples, as well as the Gaussian components are shown in Fig. 2(a). The typical samples are marked by magenta dots, whereas the outliers are marked by cyan crosses. On the other hand, the colors of the observations also indicate the values of weights $\omega_n \in [0,1]$ which are represented by cyan when $\omega_n = 0$ and by magenta when $\omega_n = 1$. The colors of samples vary smoothly from cyan to magenta as the values of $\{\omega_n\}$ approach from zero to one.

Fig. 2(b)-(f) demonstrate the learning process of proposed deterministic annealing based outlier detection method (we refer it as "DAOD" here). At the beginning, three components are randomly initialized among the samples, and the values $\{\omega_n\}$ at high temperature are almost same. As T is lowered, the components converge to the true model, and the atypical observations are gradually detected and eliminated (depicted by the smooth varying of the colors from purple to cyan). Fig. 2(e) also shows the result of conventional EM, which is marked by dashed line. It can be observed that it

fails to fit the samples due to the existing of outliers. The proposed method is also tested with different levels of trimming (1- *M/N* = 0.25, 0.35, 0.45), which is presented in Fig. 2(f). We can observe that the proposed method can still identify the clusters with higher or lower trimming level, which indicate that our algorithm is also robust to the trimming percentage. In other words, even when the prior knowledge of noise level is not consistent with the true one, our method can still give reasonable results which fit the samples appropriately.

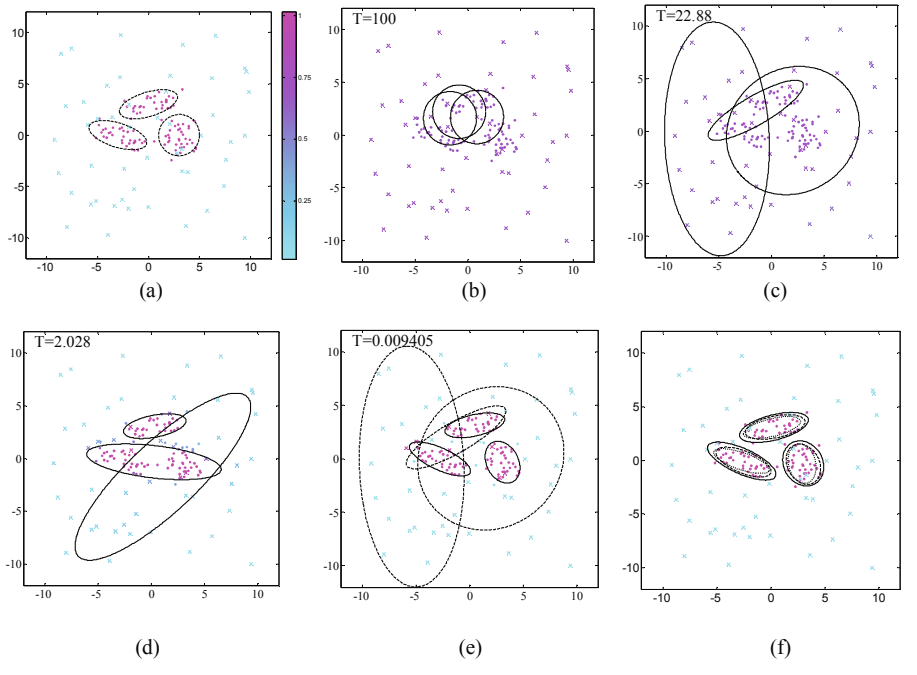

Fig. 2. Fitting Gaussian mixtures with noise: (a) the true model shown by dashed line; (b)-(d) the learning process of DAOD; (e) the results of DAOD (solid line) and the conventional EM algorithm; (f) final estimates with different trimming levels

The second example is a more complicated one since it has more components and a higher degree of overlap than the first one. This data set consists of 1000 samples from a mixture of eight two-dimensional Gaussian components with equal mixing coefficients (see also [29]), to which 250 outliers are added from a uniform distribution within [-3, 3] on each dimension, and the parameter set of each component is given by

$$\boldsymbol{\mu}_1 = [1.5,\ 0]^T\ \boldsymbol{\mu}_2 = [1,\ 1]^T\ \boldsymbol{\mu}_3 = [0,\ 1.5]^T\ \boldsymbol{\mu}_4 = [-1,\ 1]^T$$

$$\boldsymbol{\mu}_5 = [-1.5,\ 0]^T\ \boldsymbol{\mu}_6 = [-1,\ -1]^T\ \boldsymbol{\mu}_7 = [0,\ -1.5]^T\ \boldsymbol{\mu}_8 = [1,\ -1]^T$$

and

$$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_5 = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.1 \end{bmatrix} \quad \boldsymbol{\Sigma}_3 = \boldsymbol{\Sigma}_7 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.01 \end{bmatrix} \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_4 = \boldsymbol{\Sigma}_6 = \boldsymbol{\Sigma}_8 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}$$

Fig. 3 illustrates the data set, the results obtained by conventional EM, as well as DAOD with different trimming percentages. Again, it can be observed that the conventional EM cannot fit the typical observations correctly since the outliers are fitted by some components whereas some samples generated by more than one component are fitted by a single Gaussian. On the contrary, the results of our robust algorithm with different trimming percentages in Fig 3(c) indicate that our robust algorithm can locate the components correctly. The change of the trimming level only affects the estimation of covariances.

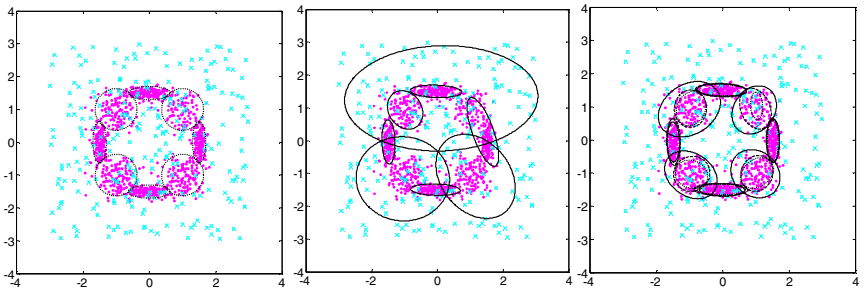

Fig. 3. Fitting an eight-components Gaussian mixture with noise (the typical samples are marked by dots, and the outliers are marked by crosses): (a) the true model presented by dashed line; (b) result of the conventional EM; (c) final estimates of DAOD with different trimming levels

To further evaluate and compare the performances of conventional EM algorithm, FAST-TLE and DAOD, we repeat the second example with different noise levels, and then compare their classification accuracies. To check the dependence of the algorithms on the initial conditions, we repeat the experiments 50 times for each noise level. In unsupervised learning scenario, labels of samples are not needed, and the classification accuracies are evaluated as below. After fitting mixtures, the samples are first partitioned into different clusters according their posterior probabilities to each component. Since the true label of each sample is known in prior, the label of each estimated cluster is assigned as the label that most samples in this cluster have. Then the sample of which the label does not agree with the cluster label is considered as misclassified, and therefore the classification accuracy can be calculated. Fig. 4 demonstrates the classification accuracies of different algorithms. Notice that the all of the algorithms have high accuracies when there is no outlier. The conventional EM algorithm is very sensitive to outliers. Both FAST-TLE and DAOD perform well when the noise level is not high. However, as the number of outliers increases, performance of FAST-TLE degrades significantly. One the other hand, DAOD can mitigate the local optima problem with FAST-TLE and therefore is more robust than the other methods.

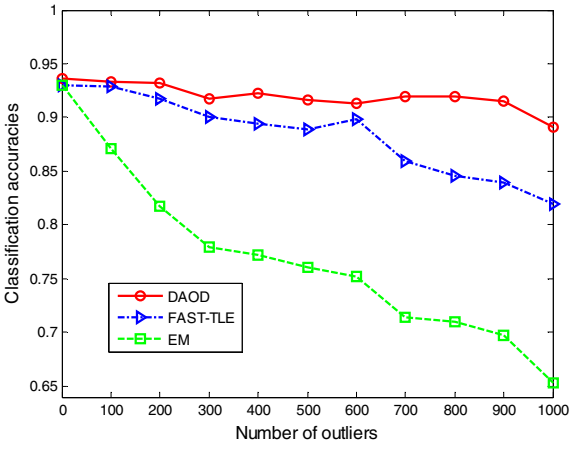

Fig. 4. Classification accuracies of various algorithms as a function number of outliers (the number of typical samples = 1000)

3.2 Real World Data Sets

We now consider the classification task for two real world data sets, i.e., Iris data set and waveform data set, which are all available from the UCI machine learning repository [24]. The Iris data set contains 150 four-dimensional samples from three classes, with each class consists of 50 samples. The waveform data set contains 5000 instances of three classes of waves, and each sample consists of 40 attributes. Therefore, we use three Gaussian components to fit each data set. The label for each component is set to dominant label of the samples, and each sample is classified according to its corresponding posterior probability.

The average classification accuracies and the standard deviations over 50 runs are demonstrated in Table I, from which it can be observed that the classification accuracy can be improved by DAOD with different trimming levels. In summary, the performance of conventional EM algorithm can be improved by gradually pruning off some samples which are located at the boundary of the components, so that the estimates will be more robust and reliable.

Table 1. Classification Accuracy (%) for Two Data Sets with Different Trimming Levels

Data Set	Proposed Robust Approach (with different trimming levels)			Conventional EM
	0.03	0.05	0.1	
Iris	94.03±4.92	95.52±2.74	94.10±4.38	93.21±4.66
Waveform	81.66±0.49	81.47±0.46	80.74±0.50	79.46±0.48

4 EEG Data Set

Finally, the proposed approach is evaluated on a more realistic application – the classification task of EEG signals. We applied our algorithm to the data set IIa from BCI competition IV [25], which consists of EEG data sets from 9 subjects. For each subject, two sessions were recorded, each of which consists of 288 trials with duration of 7s. In addition, the data set for each subject also contains some rejected trials, which are contaminated by noise or artifacts. For detailed description of this data set, see [25]. Before feature extraction, the EEG signals are filtered by 8-30Hz band pass filter. Then we applied common spatial pattern (CSP), a discriminative approach decomposing the signals into spatial patterns, to extract the features from multichannel EEG signals, and three Gaussian components are used for each class. Since the CSP is a data-driven feature extraction approach, after the elimination of noise samples, we re-train the CSP and GMM classifiers.

Fig.5 illustrates the average classification accuracies of the EEG signals of nine subjects obtained by conventional EM algorithm and the proposed robust approach with different trimming levels. The classification accuracy of the EM algorithm is not improved significantly for the original EEG data (solid line). However, when the signals are contaminated by the noise samples (rejected trials), the performance of conventional EM algorithm deteriorates obviously, whereas our proposed approach can detect and eliminate the outliers, so that more robust and reliable results (dashed line) can be obtained. We further investigate the performances of DAOD for each subject. On a whole, seven out of nine subjects benefit from DAOD. In addition, for the subjects with high classification accuracies (>80%, with fewer noise samples), the improvements are not remarkable (<2%); for the subjects with lower accuracies (<80%, with more noise samples), however, the classification accuracies of four out five subjects are improved significantly (>5%). Therefore, the proposed algorithm can successfully reduce the negative effects of EEG signals contaminated by artifacts and noise.

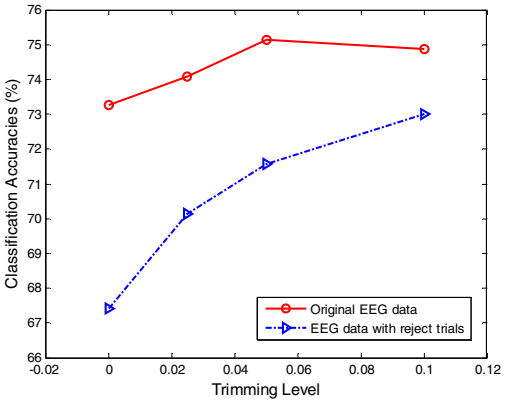

Fig. 5. The Comparison of the classification accuracies for the EEG data sets with and without rejected trials

5 Conclusion

In this paper, we proposed a DA based EM algorithm to detect outliers for mixture models. The experiments demonstrate that the performances of conventional learning approaches are significantly deteriorated due to the outliers while our method can successfully alleviate the negative effects of outliers. In addition, the proposed method can automatically prune off the EEG signals contaminated by artifacts and noise without any additional channel rejection operation (i.e., independent component analysis) or visual inspection of an expert. Since the noise is ubiquitous in EEG signals, it is necessary to prune off a small account of samples to achieve reliable result even though the noise level is unknown.

The future work will focus on the reduction on the dependence on the prior knowledge of the trimming level. It should be noted that although our method is applied to Gaussian mixtures, it can be extended to non-Gaussian cases, which will be also considered in the future work.

Acknowledgment. The authors gratefully acknowledge the support from the Macau Science and Technology Department Fund (Grant FDCT/036/2009/A) and the University of Macau Research Fund (Grants RG059/08-09S/FW/FST, RG080/09-10S/WF/FST and MYRG139 (Y1-L2)-FST11-WF).

References

1. Wolpaw, J.R., et al.: Brain–computer interface technology: a review of the first international meeting. IEEE Transactions on Rehabilitation Engineering 8(2), 164–173 (2000)
2. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-computer interface for communication and control. Clinical Neurophysiology 133(6), 767–791 (2002)
3. Bashashati, A., Fatourechi, M., Ward, R.K., Birch, G.E.: A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals. Journal of Neural Engineering 4(2), R32–R57 (2007)
4. McFarland, D.J., Anderson, C.W., Müller, K.-R., Schlogl, A., Krusienski, D.J.: BCI meeting 2005-workshop on BCI signal processing: feature extraction and translation. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2), 135–138 (2006)
5. Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain-computer interfaces. Journal of Neural Engineering 4(2), R1–R13 (2007)
6. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
7. Millán, J.R., Renkens, F., Mouriño, J., Gerstner, W.: Brain-actuated interaction. Artificial Intelligence 159, 241–259 (2004)
8. Millán, J.R.: On the need for on-line learning in brain–computer interfaces. In: Proceedings of International Joint Conference on Neural Networks, Budapest, Hungary, pp. 2877–2882 (2004)
9. Buttfield, A., Millán, J.R.: Online classifier adaptation in brain-computer interfaces. Techical Report, IDIAP–RR 06-16 (2006)

10. Sun, S., Zhang, C., Lu, N.: On the On-line Learning Algorithms for EEG Signal Classification in Brain Computer Interfaces. In: Wang, L., Jin, Y. (eds.) FSKD 2005. LNCS (LNAI), vol. 3614, pp. 638–647. Springer, Heidelberg (2005)
11. Sun, S., Zhang, C.: Learning On-line Classification via Decorrelated LMS Algorithm: Application to Brain–Computer Interfaces. In: Hoffmann, A., Motoda, H., Scheffer, T. (eds.) DS 2005. LNCS (LNAI), vol. 3735, pp. 215–226. Springer, Heidelberg (2005)
12. Schalk, G., Brunner, P., Gerhardt, L.A., Bischof, H., Wolpaw, J.R.: Brain-computer interfaces (BCIs): Detection instead of classification. Neuroscience Methods 167(1), 51–62 (2008)
13. Fazli, S., Danóczy, M., Popescu, F., Blankertz, B., Müller, K.-R.: Using Rest Class and Control Paradigms for Brain Computer Interfacing. In: Cabestany, J., Sandoval, F., Prieto, A., Corchado, J.M. (eds.) IWANN 2009. LNCS, vol. 5517, pp. 651–665. Springer, Heidelberg (2009)
14. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
15. Nguyen, D.T., Chen, L., Chan, C.K.: An outlier-aware data clustering algorithm in mixture model. In: Proceedings of 7th IEEE International Conference on Information, Communication and Signal Processing, Macau, China, pp. 1–5 (2009)
16. Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., Müller, K.-R.: Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Processing Magazine 25(1), 41–56 (2008)
17. Hadi, A.S., Luceño, A.: Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. Computational Statistics & Data Analysis 25(3), 251–272 (1997)
18. Neykov, N., Filzmoser, P., Dimova, R., Neytchev, P.: Robust fitting of mixtures using trimmed likelihood estimator. Computational Statistics & Data Analysis 52(1), 299–308 (2007)
19. Neykov, N., Müller, C.: Breakdown point and computation of trimmed likelihood estimators in generalized linear models. In: Developments in Robust Statistics, pp. 277–286. Physica-Verlag, Heidelberg (2003)
20. Rose, K., Gurewitz, E., Fox, G.C.: Statistical mechanics and phase transitions in clustering. Physical Review Letters 65(8), 945–948 (1990)
21. Rose, K.: Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. Proceedings of the IEEE 86(11), 2210–2239 (1998)
22. Grant, M., Boyd, S.: CVX: Matlab software for disciplined convex programming, version 1.21, http://cvxr.com/cvx
23. Rose, K., Gurewitz, E., Fox, G.C.: Constrained clustering as an optimization method. IEEE Transactions on Pattern Analysis and Machine Intelligence 15(8), 785–794 (1993)
24. Machine Learning Repository website, http://archive.ics.uci.edu/ml/index.html
25. BCI competition IV website, http://bbci.de/competition/iv/
26. Ueda, N., Nakano, R.: Deterministic annealing EM algorithm. Neural Networks 11(2), 271–282 (1998)
27. Zhao, Q., Miller, D.J.: A deterministic, annealing-based approach for learning and model selection in finite mixture models. In: Proceedings of 29th IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, pp. V-457–V-460 (2004)
28. Figueiredo, M.A.T., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(3), 381–396 (2002)
29. Zhang, B., Zhang, C., Yi, X.: Competitive EM algorithm for finite mixture models. Pattern Recognition 37(1), 131–144 (2004)