ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 15, No. 1, 2020, *pp*.016-021

Robust nonlinear multimodal classification of Alzheimer's disease based on GMM

Ziyue Wang[1]
School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China
(Received September 07, 2019, accepted October 26, 2019)

Abstract: Accurate diagnosis of Alzheimer's disease (AD) and its prodromal stage mild cognitive impairment (MCI) is very important for patients and clinicians. There are many useful medical data have been discovered to be remarkable for diagnosis i.e., structural MR imaging (MRI), functional imaging (e.g., FDG-PET and FIB-PET). Multimodal classification model is needed to combine these biomarkers to improve the diagnose performance. Some methods have been proposed such as linear mixed kernel, combined embedding and nonlinear graph fusion. These methods have efficiently employed the multimodal data, but they ignore the influence of noise and outliers. Noise is easily generated in image analysis and measurement. To enhance robustness, mixture distributions were applied in nonlinear regression models. Gaussian mixture model is successfully applied in many domains. In this paper, we generalize nonlinear multimodal classification model based on GMM. The performance on real dataset: 22 AD, 23 MCI and 25 NC (health) is comparable to other methods.

Keywords: Robust nonlinear regression, Outlier, Kernel method, Classification.

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia in elderly people. AD greatly affects the cognitive ability of the elderly [1]. Thus, it is important to diagnose AD as soon as possible from its early stage mild cognitive impairment MCI. In the clinic, many medical images and biological indicators are used for diagnosis. Such as MRI (MR image) [2], functional imaging (FDG-PET, FIB-PET) [3] and quantification of specific proteins measured through CSF [2].

Different biomarkers can contain different feature of AD patient, thus may provide complementary information for diagnosis [4-6]. In [5, 7], linear mixed kernel is proposed independently. Paper [5] learn the kernel weight by grid search while paper [7] take the kernel weight as model parameter and learn it by optimization. Similarities from multiple modalities are combined to generate an embedding, which contain information of multimodal data [6]. In paper [4], similarity matrix for classification is calculated by nonlinear graph fusion. In this paper, kernel method is also used for multimodal data, and the construction of the combined kernel matrix is the same as the mixed kernel of paper [5].

After the construction of the mixed kernel, which contains sample information completely, efficient classification is needed. There are many classification models have been proposed such as logistic regression, k-nearest neighbor, naïve bayes, decision tree, SVM [8-10] and so on. However, most of those classification models do not model the noise directly except support vector machine. Support vector machine model the noises and outliers with the slack variable. In SVM, the input data is mapped into a higher dimensional space to make it separable. SVM can solve two-class classification, and the goal is to maximize the decision bound. This method is totally influenced by the support vectors on the decision bound, if most of those support vectors are polluted by noises, the model will be not proper enough. Therefore, the slack variable is proposed to make the decision bound more robust. Kernel method is improved to deal with the nonlinear case.

Based on the traditional SVM, Least-Squares SVM (LSSVM) [11] is proposed. The LSSVM changes the equality constraint in SVM to the inequality constraint. As a result, the convex quadratic programming is replaced to convex linear problem. In Least-Squares SVM, the slack variables are proportional to the errors.

Mixture models are successfully applied in many domains due to their excellent robustness. In paper [12, 13], mixture of t and skew normal distribution is applied separately to fit the noise term in the linear

[1] Corresponding author. E-mail address: 824831789@qq.com.

regression model. In [14, 15], Gaussian mixture models (GMMs) based classification models are applied in medical research. Paper [16] uses the Gaussian mixture models (GMMs) for multiple limb motion classification using continuous myoelectric signals. Besides, mixture model applied in machine learning [17].

In real world, the data is usually polluted by outliers and heavy-tailed noises, the slack variable in Least-Squares SVM can't be well characterized. In this paper, we develop a nonlinear classification model while the feature of noise is fitted by Gaussian mixture model (GMM). The linear mixed kernel method is employed which contains the multimodal data. In order to get the optimal parameter, EM algorithm and Lagrange multiplier method are applied. The experiment results are comparable to other multimodal-based classification methods.

2. Methodology

2.1 Nonlinear classification model

Given the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_i \in R^d$ is the input data , $y_i \in \{-1,\ 1\}$ is the label. The objective function of support vector machine is:

$$f(x) = sign[\sum_{i=1}^n \alpha_i k(x_i, x) + b] \tag{1}$$

where α_i is the parameter of Lagrange multiplier method, $k(x_i, x)$ is the kernel function, b is the bias.

Assuming that

$$\begin{cases} \boldsymbol{\omega}^T \phi(\boldsymbol{x}_i) + b = 1 - e_i \\ \boldsymbol{\omega}^T \phi(\boldsymbol{x}_i) + b = -1 + e_i \end{cases}$$

then, we have

$$(\boldsymbol{\omega}^T \phi(\boldsymbol{x}_i) + b) y_i = 1 - e_i, \quad i = 1,2,\cdots,n$$

where $\phi(\cdot)$is the map function, e_i is the error.

The objective function of SVM is:

$$\begin{aligned} &\min_{w,e} \ \frac{1}{2}\|\boldsymbol{\omega}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{e}\|_2^2 \\ &s.t. \quad \boldsymbol{G}^T\boldsymbol{\omega} + b\boldsymbol{y} = \boldsymbol{1}_n - \boldsymbol{e} \end{aligned} \tag{2}$$

where $\boldsymbol{G} = (y_1\phi(x_1), y_2\phi(x_2), \dots, y_n\phi(x_n)) \in R^d \times n$, $\boldsymbol{e} = (e_1, e_2, \dots, e_n)$ is the error, γ is a regularized parameter.

2.2 GMM based nonlinear classification model

Gaussian mixture model:

$$p(e) = \sum_{k=1}^K \pi_k N(e|0, \sigma_k^2) \tag{3}$$

where K is the number of independent Gaussian distribution in GMM model. $N(e|0, \sigma_k^2)$ is the Gaussian distribution with zero mean, variance σ_k^2, π_k is the weight coefficient that satisfied: $\sum_{k=1}^K \pi_k = 1,\ \pi_k \geq 0$.

Theoretically, we need to define the form of the map function $\phi(x)$ in advance. However, this will increase the number of coefficient and computation complexity, in the same time, choosing a mapping function is complicated. Similar to LS-SVM, we will use the Lagrange multiplier method in the optimize step. So the map function always appears as $\phi(x)^T\phi(x)$. Therefore, we can introduce kernel function $k(x, y) = \phi(x)^T\phi(y)$. In this paper, RBF kernel is employed:

$$k(x, y) = exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{4}$$

The optimal values of parameter can be obtained by maximum likelihood estimation, the likelihood function of e can be expressed as:

$$p(e|\Theta) = \prod_{i=1}^n p(e_i|\Theta) = \prod_{i=1}^n \sum_{k=1}^K \pi_k N(0, \sigma_k^2) \tag{5}$$

where Θ is the parameter set. Then the log-likelihood function is calculated as:

$$L(e|\Theta) = \sum_{i=1}^n log\, p\,(e_i|\Theta) = \sum_{i=1}^n (log \sum_{k=1}^K \pi_k N(0, \sigma_k^2)) \tag{6}$$

Due to the complex expression of log-likelihood function, it is difficult to calculate directly. The EM algorithm is an efficient algorithm to solve such problems.

In order to simplify the solution process, we introduce $\boldsymbol{Z} = (z_1, z_2, \dots, z_n)^T$, where $z_i = (z_{i1}, z_{i2}, \dots, z_{iK})$ is an indicator vector, if e_i comes from the jth component, then $z_{ij} = 1$ the other elements of z_i are 0. So $\sum_{k=1}^K z_{ik} = 1$, $\sum_{i=1}^n \sum_{k=1}^K z_{ik} = 1$.

z_i obeys multi-point distribution:

$$z_i \sim MN(1, \boldsymbol{\pi}),\ \boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)$$

There exist latent variable $u_i, i = 1,2, \dots, n$ that satisfies:

$$p(e_i | z_{ik} = 1) = N(0, \sigma_k^2) \tag{7}$$

Then the parameter set is $\chi = (e, z_1, \dots, z_n)$.The log-likelihood function is calculated as

$$L(\chi|\Theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[log\ \ \pi_k - \frac{1}{2} log\ \ \sigma_k^2 - \frac{1}{2\sigma_k^2} \cdot e_i^2 \right]. \tag{8}$$

2.3 Parameter estimation

In the E step of EM algorithm, based on the obtained parameter $\Theta$and the observe data calculated in the last step, $Q(\Theta^t|\Theta^{t-1})$ can be obtained by computing the conditional expectation of $L(\chi|\Theta)$ with respect to z_{ik}.

Based on z_{ik} obeys multi-point distribution, its expectation can be obtained as follows:

$$\gamma_{\text{ik}} = E(z_{ik}|e_i) = \frac{\pi_k \cdot N(e_i|0,\sigma_k^2)}{\sum_{k=1}^{K} \pi_k \cdot N(e_i|0,\sigma_k^2)}. \tag{9}$$

The Q function $Q(\Theta^t|\Theta^{t-1})$ is:

$$Q(\Theta^t|\Theta^{t-1}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left[log\ \ \pi_k - \frac{1}{2} log\ \ \sigma_k^2 - \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{e_i^2}{2\sigma_k^2} \right] \tag{10}$$

In the M step of EM algorithm, we will update the parameter space by maximizing the Q function.

Update π

Take $\sum_{k=1}^{K} \pi_k = 1,\ \pi_k \geq 0$ as a constraint, employing the Lagrange multiplier method, the update formula of π is:

$$\pi_k = \frac{\sum_{i=1}^{n} \gamma_{ik}}{n}, k = 1,2, \dots, K. \tag{11}$$

Update Σ

$$\sigma_k^2 = \frac{\sum_{i=1}^{n} \gamma_{ik} \cdot e_i^2}{\sum_{i=1}^{n} \gamma_{ik}} \tag{12}$$

Update e

e should be update by maximize the following:

$$\begin{aligned} J_{w,b} &= -\sum_{\text{i=1}}^{\text{n}} \sum_{k=1}^{K} \gamma_{ik} \cdot \frac{e_i^2}{2\sigma_k^2} \\ &= -\sum_{\text{i=1}}^{\text{n}} (\sum_{k=1}^{K} \frac{\gamma_{ik}}{2\sigma_k^2})(1 - \mathbf{G}_i^T \mathbf{w} - by_i)^2 \\ &= -\left\| \boldsymbol{\lambda} \otimes (\mathbf{1}_n - \mathbf{G}^T \mathbf{w} - b\mathbf{y}) \right\|_2^2 \end{aligned} \tag{13}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_n)^T$ is weight vector, $\lambda_i = \sqrt{\sum_{k=1}^{K} \frac{\gamma_{ik}}{2\sigma_k^2}}$, $\otimes$ means the Hadamard product. $\boldsymbol{w}$ and b should be obtained by maximizing $J_{w,b}$. It can be transformed into

$$\begin{aligned} &arg \min_{w,b,e}\ \ \|\boldsymbol{\lambda} \otimes \boldsymbol{e}\|_2^2 \\ &s.t.\quad \boldsymbol{e} = \mathbf{1}_n - \boldsymbol{G}^T \boldsymbol{w} - b\boldsymbol{y} \end{aligned} \tag{14}$$

Regularization coefficient should be add to avoid overfitting:

$$\begin{aligned} &arg \min_{w,b,e}\ \frac{\beta}{2} \|\boldsymbol{\lambda} \otimes \boldsymbol{e}\|_2^2 + \frac{1}{2} \|w\|_2^2 \\ &s.t.\quad \boldsymbol{e} = \mathbf{1}_n - \boldsymbol{G}^T \boldsymbol{w} - b\boldsymbol{y} \end{aligned} \tag{15}$$

Similar to LS-SVM , we employ Lagrange multiplier method, the Lagrange function is:

$$L(\boldsymbol{w}, \boldsymbol{e}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\boldsymbol{w}\|_2^2 + \frac{\beta}{2} \|\boldsymbol{\lambda} \otimes \boldsymbol{e}\|_2^2 - \boldsymbol{\alpha}^T (\boldsymbol{e} - \mathbf{1}_n + \boldsymbol{G}^T \boldsymbol{w} + b\boldsymbol{y}). \tag{16}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$ means Lagrange multiplier, according to KKT conditions:

$$\begin{cases} \frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \boldsymbol{w} = \boldsymbol{G}\boldsymbol{\alpha} \\ \frac{\partial L}{\partial b} = 0 \quad \Rightarrow \boldsymbol{\alpha}^T \boldsymbol{y} = 0 \\ \frac{\partial L}{\partial \boldsymbol{e}} = 0 \quad \Rightarrow diag(\boldsymbol{e}) = \beta^{-1} diag(\boldsymbol{\lambda})^{-2} diag(\boldsymbol{\alpha}) \\ \frac{\partial L}{\partial \boldsymbol{\alpha}} = 0 \quad \Rightarrow \boldsymbol{G}^T \boldsymbol{w} + b\boldsymbol{y} - \mathbf{1}_n + \boldsymbol{e} = \mathbf{0}_n \end{cases} \tag{17}$$

DOI https://doi.org/2024-JICS-22393 | Generated on 2025-04-20 20:32:48

By eliminating w and e ,solution is given by the following set of linear equation:

$$\begin{bmatrix} \mathbf{0} & \boldsymbol{y}^T \\ \boldsymbol{y} & \boldsymbol{H} \end{bmatrix} \times \begin{bmatrix} b \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{1}_n \end{bmatrix} \tag{18}$$

where $\boldsymbol{H} = \boldsymbol{K} + \beta^{-1} diag(\boldsymbol{\lambda})^{-2}, K = G^T G$ is kernel function and $\boldsymbol{K}(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x})^T \phi(\boldsymbol{y})$. Let the solution be $\hat{b}, \hat{\boldsymbol{\alpha}}$, then the classification model is :

$$f(\boldsymbol{x}) = sign(\sum_{i=1}^{n} \hat{\alpha}_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \hat{b}) \tag{19}$$

Then e can be updated by the following equation:

$$\mathbf{e} = \mathbf{1}_n - \boldsymbol{y}\boldsymbol{y}^T \otimes K(\boldsymbol{x}, \boldsymbol{x})\boldsymbol{\alpha} - b\boldsymbol{y} \tag{20}$$

2.4 Kernel based multimodal classification

It is easy to see in Eq.(18),(19),(20), train data is always in the kernel function. So the multimodal classification model can be construct if we turn $K(\boldsymbol{x}, \boldsymbol{y})$ into:

$$K_{mixed}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{m} K(\boldsymbol{x}, \boldsymbol{y}) \tag{21}$$

where m is the number of modal.

The construct of the kernel based multimodal classification can be seen in fig1:

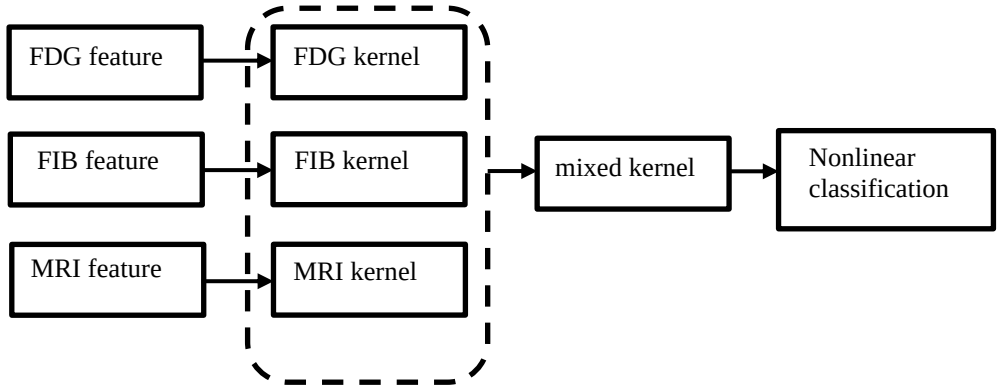

Fig1: Overview of the proposed framework

3. Experiments and results

3.1 Subject

FDG-PET contains 90 features of 90 regions, FIB-PET contains 90 features of 90 regions, MRI contains 90 features from 90 regions. The description of 70 subject is in Table 1.

Table 1: Description of Subjects

	AD	MCI	NC(health)
Number	22	23	25

3.2 Implementation details

For all the experiments data, the input and target variables are normalized into the interval [0, 1]. The Gaussian distribution number of GMM $K = 2$. The weights in the multi-kernel method are learned based on the training samples, through a grid search using the range from 0 to 1 at a step size of 0.1. The regularize parameter of the classification method β and the kernel parameter σ^2 are learned from the set $\{2^i | i =$ -16,-6, … ,0, … ,6,16$\}$. Because the kernel weight parameter is dependent by the train data, while the regularize parameter and kernel parameter are dependent by the model, we train them separately. Firstly, we learn the kernel weight parameter with fixed $\beta = 1, \sigma^2 = 1$. Then, we train the model with the optimal kernel weight parameter.

The model multiple kernel learning(MKL) [5], combined embedding(CE) [6], nonlinear graph fusion (NGF) [7] are used for compare. The classification is implement on AD vs. NC (47 samples) and MCI vs. NC (48 samples). They all randomly selected 30 samples for the training set and the rest for the test set. Since the training sample set is not rich enough, we have adopted Data Augmentation to prevent overfitting

by duplicate the training set 5 times. Experiments results are get by 5 fold cross-validation. The results are in Table2:

Table2: Comparison of results based on multi-modality classification

Classification	Metrics(%)	MKL[5]	CE[6]	NGF[4]	NLC-GMM
AD VS. NC	Accuracy	91.8	92.8	95.2	96.4
	Sensitivity	87.6	89.4	93.3	95.2
	Specificity	95.8	96.9	97.2	97.6
MCI VS. NC	Accuracy	70.0	74.4	75.7	78.8
	Sensitivity	72.4	69.6	77.6	77.3
	Specificity	68.6	80.0	74.1	79.6

It is easy to see that the proposed GMM based nonlinear classification (NLC-GMM) is robust to the noise and outlier contained in the train data. As in the table 2, the accuracy, sensitivity and specificity is significantly higher than the linear kernel based MKL [5]. Our method is also better than CE [6] and NGF [7], which explains the effectiveness of GMM noise modeling.

The output of the Eq. (19) is:

$$\text{out}put = \sum_{i=1}^{n} \hat{\alpha}_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \hat{b} \tag{22}$$

The output is real number, which can be mapped into interval [0,1] by the sigmoid function:

$$f(x) = \frac{1}{1+exp(-x)} \tag{23}$$

$f(output)$ is the probability that the sample belongs to AD (AD vs. NC) or MCI (MCI vs. NC).

After getting the probability, we the ROC curve is in the Fig2:

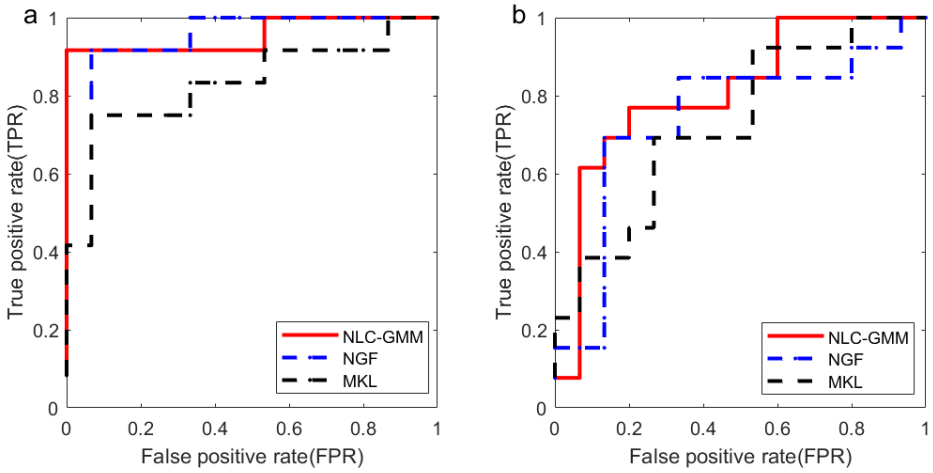

Fig2: Roc curves of different methods. (a) AD vs. NC, (b) MCI vs. NC

4. Conclusion and future work

In this work, we focus on the performance of classification model on the data, which is contaminated by unknown noise and outliers. By introducing the powerful Gaussian mixture model (GMM), and the linear mixed kernel [5], the performance and robustness of the nonlinear classification model is significantly improved. This result shows that noise is inevitable during the data acquisition process. This phenomenon suggests that we should build a model that can automatically identify noise or select features before model training.

In medical image data, a patient usually has many features, so it is likely to contain useless and redundant information. Therefore, the feature selection is necessary in this situation. Feature selection method based on t-test statistics is employed in [5].

The information matrix based on sample similarity is proposed by [4, 6]. The performance of this method reveals that the information matrix carries more information than the RBF kernel. Therefore, it is interesting to replace the linear mixed RBF kernel to the information. In addition, many robust models based on mixed distributions have been proposed such as Mixture of t distribution [18] and scale mixture of skew-normal [13]. The performance of those mixture models on the different dataset is worth studying.

References

[1] J. Cr, D. S. Knopman, W. J. Jagust, L. Shaw, et al. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade [J]. Lancet Neurology, 2010, 9(1): 4-5.

[2] A. M. Fjell, K. Walhovd, C. Fennema, et al. CSF Biomarkers in Prediction of Cerebral and Clinical Change in Mild Cognitive Impairment and Alzheimer's Disease [J]. The Journal of Neuroscience, 2010, 30(6): 2088-2101.

[3] J. Langbaum, K. Chen, W. Lee, et al. Categorical and correlational analyses of baseline fluorodeoxyglucose positron emission tomography images from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [J]. NeuroImage, 2009, 45(4): 1107-1116.

[4] T. Tong, K. Gray, Q. Gao, et al. Multi-Modal Classification of Alzheimer's Disease Using Nonlinear Graph Fusion [J]. Pattern Recognition, 2016, 63: 171-181.

[5] D. Zhang, Y. Wang, L. Zhou, et al. Multimodal classification of Alzheimer\"s disease and mild cognitive impairment [J]. Neuroimage, 2011, 55(3): 856-867.

[6] K. R. Gray, P. Aljabar, R. A. Heckemann, et al. Random forest-based similarity measures for multi-modal classification of Alzheimer disease [J]. Neuroimage, 2013, 65: 167-175.

[7] C. Hinrichs, V. Singh, G. Xu, et al. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population [J]. Neuroimage, 2011, 55(2): 574-589.

[8] V. Vapnik. “The nature of statistical learning theory,” Springer-Verlag, New-York, 1995.

[9] V. Vapnik. “Statistical learning theory,” John Wiley, New-York, 1998.

[10] V. Vapnik. “The support vector method of function estimation,” In Nonlinear Modeling: advanced black-box techniques, Suykens J.A.K., Vandewalle J. (Eds.), Kluwer Academic Publishers, Boston, pp. 55-85, 1998.

[11] J. A. Suykens, J. Vandewalle. Least Squares Support Vector Machine Classifiers [J]. Neural Processing Letters, 1999, 9(3): 293-300.

[12] W. Yao, Y. Wei, C. Yu. Robust mixture regression using the t-distribution [J]. Computational Statistics and Data Analysis, 2014, 71(3): 116-127.

[13] C. B. Zeller, C. R. Cabral, H. Vctor. Robust mixture regression modeling based on scale mixtures of skew-normal distributions [J]. TEST, 2016, 25(2): 375-396.

[14] E. M. Thomas, A. Temko, G. Lightbody, et al. A Gaussian mixture model based statistical classification system for neonatal seizure detection[C]// Machine Learning for Signal Processing, 2009. IEEE International Workshop on IEEE.

[15] S. Khanmohammadi, C. A. Chou. A Gaussian Mixture Model Based Discretization Algorithm for Associative Classification of Medical Data [J]. Expert Systems with Applications, 2016, 58.

[16] N. Dilokthanakul, P. A. Mediano, M. Garnelo, et al. Deep unsupervised clustering with Gaussian mixture variational autoencoders. arXiv:1611.02648, 2016.

[17] Y. Huang, K. B. Englehart, B. Hudgins, et al. A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses [J]. IEEE Trans Biomed Eng, 2005, 52(11): 1801-1811.

[18] G. Galimberti, G. Soffritti. A multivariate linear regression analysis using finite mixtures of t distributions [J]. Computational Statistics & Data Analysis, 2014, 71(3): 138-150.