Feature Selection for Diffuse Lung Disease using Exchange Markov Chain Monte-Carlo Method

Makoto KOIWAI¹, Nodoka IIDA¹, Hayaru SHOUNO¹, Shoji KIDO²

¹ Graduate School of Informatics and Engineering, University of Electro-Communications, Chofugaoka 1-5-1, Chofu, JAPAN
² Graduate School of Medicine, Yamaguchi University, Tokiwadai 2-16-1, Ube, JAPAN

Abstract—Diffuse lung disease (DLD) in high resolution computed tomography images show a lot of variations even in the same class, and this variations make difficulty in diagnosis. In this study, we treat a effective feature selection problem for this DLD pattern classification using machine learning approach. In order to obtain the best feature selection for classification, we should search whole combination of features, which requires exponential order calculation cost. Recently, Nagata et al. proposed an application of Exchange Markov Chain Monte Carlo (ExMCMC) method for this problem, and suggested that they reveals hidden feature structures for classification. Thus, we tried their method to select the effective feature combination for each DLD classification from 39 types of features, which are obtained from typical texture analysis method in the image processing. As the result, we obtained the effective feature combination candidates for each DLD classification problem.

Keywords: Feature Selection, Medical Image, replica exchange MCMC

1. Introduction

In this research, we apply a feature selection method to classify the diffuse lung disease (DLD). The DLD is a kind of inflaming which is spread in the wide area of the lung. In the last stage of the DLD, the disease site lose function of lung and the patient becomes hard to recover, so that early detection of the DLD is desired[1][2]. To diagnose the DLD, the high resolution computed tomography (HRCT) images are regarded as the effective detection of DLD, because diagnosing physician can diagnose the spreading site of the candidate area from several directions. However, there exists large varieties of the DLD patterns on the HRCT image. Thus, the early detection influenced to the skill of physician who should diagnose whole lung volumes, which has over hundreds HRCT slice images. Moreover, the introducing the second opinion system increase the burden of the physician. So the computer aided diagnosis (CAD) system for DLD is desired to construct. In order to apply this requirement, many researchers introduce pattern classification technology into the DLD diagnosis[3][4][5][6].

In the field of machine learning, the pattern classification technology consists of both feature extraction part and classifier part[7]. In these decades, several classifiers, such like support vector machine (SVM), logistic regression, Bayes method and so on, are discussed. However the discussion about feature extraction and selection method looks not enough. So, we focus on the feature selection method to find a good feature set for classification for DLD patterns. In the previous works, Sugata et al. proposed a set of texture features for DLD pattern representation and apply it for classification with Naive Bayes method[1]. Wada and Hayakawa applied semi-supervised learning method for this feature representation[2][8]. In their research, they pointed out that using the full set of feature representation makes worse classification rather than that of some selected features. Wada et al. uses only about 4 selected feature for their experiment. The excess feature representation makes classification performance worse. This phenomenon is known as “curse of dimensionality”, so that, the feature selection is important factor for classification performance essentially. The most rigid method for feature selection is using a Brute-force style method. Ichikawa et al. applied the exhaustive search of features for classifying attention deficit hyper-activity disorder (ADHD) from electroencephalogram (EEG)[9]. Unfortunately, the feature selection method is a kind of combinatorial problem, which requires exponential order calculation for exact solution search, which is sometimes called Brute-force search.

Therefore, the larger the number of whole feature set becomes, the more difficult the feature selection becomes. For this feature selection problem, Nagata et al. applied a Markov Chain Monte Carlo (MCMC) sampling with replica exchange system[10][11]. Hereinafter, we call that MCMC method as ExMCMC (replica Exchange MCMC) method. The ExMCMC is known as a powerful sampling method,
Fig. 1: Measuring method for generalization error using cross validation. Cross validation divides dataset into two parts called “training” and “test” samples. The training set is applied for constructing the classifier, and the test is for evaluation.

![Cross Validation Diagram](image1.png)

**CV score = Average Accuracy for Test Patterns**

Fig. 2: Schematic diagram of ES-SVM method with MCMC. Regarding the CV score as a kind of cost function \( H(s) \), our purpose is to find better solution candidates \( \{s\} \), which provides low \( H(s) \)

which runs the parallel MCMC in several relaxed conditions, and exchange states between parallelized run when several conditions are satisfied. Nagata et al. introduces the ExMCMC method for detecting information carrier neuron in the brain[10]. Thus, in this research, we apply ExMCMC based feature selection method for texture feature representation for DLD patterns, and evaluate the feature for classification.

### 2. Method

In this section, we explain feature selection method. The key idea comes from feature selection with exhaustive search with SVM [9][10], and ExMCMC method for sample enumeration.

2.1 Feature extraction with Exhaustive Search SVM method

The most sure method for feature selection is to apply exhaustive search which means “Brute-force” search. Here let us consider the following situation, that is, we have \( D \) features in the observation and want to find the most effective features combination for classifying. Before constructing a classifier, we must choose a feature set for the input of the classifier. When we choose a feature set with some method, we can evaluate the performance of the classifier with cross validation (CV) method. Cross validation is a kind of measure for generalization error[12]. Fig. 1 shows a concept for cross validation method. When we divide \( K \) sub-dataset, we can choose one subset as a ‘test set’ and the other as a ‘training set’. We train the classifier with training set and evaluate the classification performance with test set. Thus, we can regard the classification performance for the novel input pattern. But it contains some arbitrary selection for the test set, so that we evaluate whole combination of the training set and test set. The cross validation score means the average of the whole combination of the classification performance. This is the concept of the K-fold CV method.

The most sure way for feature selection is to search the feature set, which shows the best CV score, from the whole combination of feature sets. Ichikawa et al. search the most effective for classifying ADHD disease from the 24 channel EEG[9]. Also, Nagata et al. introduce this feature selection method for picking up a neuron set in the brain to determine the information carrier for face recognition.

Unfortunately, the calculation cost for search whole combinations becomes \( O(2^D) \), so that this method requires exponential order calculation cost. The more convenient way is to introduce some sparse prior, such like \( L1 \) prior, automatic relevant determination (ARD), and so on. However, Nagata pointed out the results of the \( L1 \) sparse logistic regression and the one with ARD showed different results. From the viewpoint of the accuracy of the feature set, we should not discard the exhaustive search method if we can calculate whole combination.

We introduce linear SVM as the classifier in this research. Thus, hereinafter, we call the exhaustive search method as “ES-SVM” method. The linear SVM is a simple linear classifier which divide input space into a hyper-plane, which is called decision boundary or discrimination plane, characterized normal vector \( \mathbf{w} \) and interception \( b \). The decision boundary is formulated as \( y(x) = \mathbf{w}^T \mathbf{x} + b = 0 \). When a novel input \( \mathbf{x}_{novel} \) is input to the system, the SVM evaluate whether the novel input is included in the target class or not with the value of \( y(\mathbf{x}_{novel}) \). If \( y(\mathbf{x}_{novel}) > 0 \), the novel input...
Our purpose is to find some desirable transition. The driving force of the MCMC method is to overcome the transition problem[11][10]. We introduce temperature parameter $T > 0$ and its inverse $\beta = 1/T$. Considering the probability with inverse temperature $\beta$ of probability, we can re-define the probability with weight by inverse temperature $p(s) \propto \exp(-\beta H(s))$. The temperature $\beta = 1$ means our original cost function. When $\beta$ becomes small, the efficacy of the cost function $H(s)$ also becomes small. So the landscape of the weighted cost function $\beta H(s)$ becomes smooth. Fig.3 shows the concept of the replica exchange MCMC method. We prepare $L$ parallel replicated MCMC system, and we run each MCMC with different temperature $T_i$. After several MCS, we exchange several replica states. As the result, we can obtain sample from wide spreading multiple peak distribution via low temperature Markov-chain transitions. The procedure for the ExMCMC is summarized as following:

1) Prepare $M$ replicated systems, and assign appropriate inverse temperature $0 < \beta_0 < \beta_1 < \cdots < \beta_{M-1} = 1$. Denoting each system status variable as $s_m$, where $m$ means the index of system.
2) Carrying out several MCSs under the probability of $p(s_m) \propto \exp(-\beta_m H(s_m))$ for $m$th system. Now, we describe the exchange timing as $\tau$.
3) Select one temperature site denoted as $j$.  

The replica exchange MCMC method is to overcome the transition problem[11][10]. We introduce temperature parameter $T > 0$ and its inverse $\beta = 1/T$. Considering the probability with inverse temperature $\beta$ of probability, we can re-define the probability with weight by inverse temperature $p(s) \propto \exp(-\beta H(s))$. The temperature $\beta = 1$ means our original cost function. When $\beta$ becomes small, the efficacy of the cost function $H(s)$ also becomes small. So the landscape of the weighted cost function $\beta H(s)$ becomes smooth. Fig.3 shows the concept of the replica exchange method. We prepare $L$ parallel replicated MCMC system, and we run each MCMC with different temperature $T_i$. After several MCS, we exchange several replica states. As the result, we can obtain sample from wide spreading multiple peak distribution via low temperature Markov-chain transitions. The procedure for the ExMCMC is summarized as following:

1) Prepare $M$ replicated systems, and assign appropriate inverse temperature $0 < \beta_0 < \beta_1 < \cdots < \beta_{M-1} = 1$. Denoting each system status variable as $s_m$, where $m$ means the index of system.
2) Carrying out several MCSs under the probability of $p(s_m) \propto \exp(-\beta_m H(s_m))$ for $m$th system. Now, we describe the exchange timing as $\tau$.
3) Select one temperature site denoted as $j$.  

2.2 Sample Enumeration with replica exchange Markov-chain Monte Carlo method

The problem for the exhaustive search for feature selection is the exponential calculation cost. We introduce a kind of Markov chain Monte-Carlo (MCMC) method for the enumeration of the feature set. Fig.2 shows the schematic diagram for the ES-SVM with MCMC method. In the figure, we have $D$ feature candidates. Thus, we represent a feature set as state variable $s = \{s_d\}_{d=1}^{D}$ in which each element $s_d$ has binary value $s_d \in \{0, 1\}$ that means the selected feature or not. Our purpose is to enumerate the $s$, which minimize the CV score in the previous section. Thus, we introduce the cost function $H(s)$ as the CV score. Using optimization method for minimization $H(s)$, we can only find one vector set of $s$. Our purpose is to find some candidates set of $s$, so that, enumeration method is better rather than the optimization. In this research we adopt a Markov chain Monte Carlo (MCMC) sampling method for enumeration. The procedure for the MCMC is summarized as following:

1) Select one site $s_i^{(t)}$ in the state vector $s^{(t)}$ where $t$ means the time index.
2) Prepare a candidate vector $s^*$ in which only $s_i^{(t)}$ is inverted from the vector $s^{(t)}$.
3) Calculate the costs $H(s^{(t)})$ and $H(s^*)$ and evaluate the probability

$$r = \min \left(1, \frac{\exp(-H(s^*))}{\exp(-H(s^{(t)}))} \right). \quad (1)$$

4) Generate an unit random value $u \in [0, 1]$, and compare $u$ with the $r$. If $u < r$ then accept the state $s^*$ as a new state $s^{(t+1)}$, and the other case the state is hold as $s^{(t+1)} = s^{(t)}$.
5) Goto the 1st step while $t$ satisfies the iteration limit

This method is known as Metropolis-Hasting (MH) method[14], and hereinafter we call this successive procedures as Monte Carlo step (MCS). Using the MCMC method, we can obtain samples $\{s^{(t)}\}$ which obeys the probability $p(s) \propto \exp(-H(s))$.

The MH method is a strong method for enumerating, however, it requires long calculation time to sample from wide spreading multiple peak distribution. For transition from a peak to another, there exists low probability region in any transition paths. The driving force of the MCMC depends on the odds ratio of the pre- and post- state in eq.(1). So that, too much low probability region prohibits desirable transition.

2.2 Sample Enumeration with replica exchange Markov-chain Monte Carlo method

The problem for the exhaustive search for feature selection is the exponential calculation cost. We introduce a kind of Markov chain Monte-Carlo (MCMC) method for the enumeration of the feature set. Fig.2 shows the schematic diagram for the ES-SVM with MCMC method. In the figure, we have $D$ feature candidates. Thus, we represent a feature set as state variable $s = \{s_d\}_{d=1}^{D}$ in which each element $s_d$ has binary value $s_d \in \{0, 1\}$ that means the selected feature or not. Our purpose is to enumerate the $s$, which minimize the CV score in the previous section. Thus, we introduce the cost function $H(s)$ as the CV score. Using optimization method for minimization $H(s)$, we can only find one vector set of $s$. Our purpose is to find some candidates set of $s$, so that, enumeration method is better rather than the optimization. In this research we adopt a Markov chain Monte Carlo (MCMC) sampling method for enumeration. The procedure for the MCMC is summarized as following:

1) Select one site $s_i^{(t)}$ in the state vector $s^{(t)}$ where $t$ means the time index.
2) Prepare a candidate vector $s^*$ in which only $s_i^{(t)}$ is inverted from the vector $s^{(t)}$.
3) Calculate the costs $H(s^{(t)})$ and $H(s^*)$ and evaluate the probability

$$r = \min \left(1, \frac{\exp(-H(s^*))}{\exp(-H(s^{(t)}))} \right). \quad (1)$$

4) Generate an unit random value $u \in [0, 1]$, and compare $u$ with the $r$. If $u < r$ then accept the state $s^*$ as a new state $s^{(t+1)}$, and the other case the state is hold as $s^{(t+1)} = s^{(t)}$.
5) Goto the 1st step while $t$ satisfies the iteration limit

This method is known as Metropolis-Hasting (MH) method[14], and hereinafter we call this successive procedures as Monte Carlo step (MCS). Using the MCMC method, we can obtain samples $\{s^{(t)}\}$ which obeys the probability $p(s) \propto \exp(-H(s))$.

The MH method is a strong method for enumerating, however, it requires long calculation time to sample from wide spreading multiple peak distribution. For transition from a peak to another, there exists low probability region in any transition paths. The driving force of the MCMC depends on the odds ratio of the pre- and post- state in eq.(1). So that, too much low probability region prohibits desirable transition.
3.1 Materials

In this research, we prepare 360 labeled images. Each class has following number of images: Consolidation(CON):38, Ground-Grass-Opacity(GGO):76, Honeycomb(HCM):49, Reticular(RET):37, Emphysema(EMP):54, Nodular(NOD):48, and Normal(NOR):58 cases. We assume the 32 × 32 [pixels] ROIs, and each ROI is segmented under the direction of a physician, and diagnosed by 3 physicians.

The acquisition parameters of those HRCT images are as follows: Each images are obtained from Toshiba “Aquilion 16” imaging device. Each slice image consists of 512 × 512 pixels, and pixel size corresponds to 0.546 ~ 0.826 [mm], slice thickness are 1 [mm]. The number of patients is 69 males and 42 females with age 66.3 ± 13.4. The number of normal donor is 4 males and 2 females with age 44.3 ± 10.3. The origin of these image data is provided Tokushima University Hospital. Fig.4 shows segmented images of typical examples of each disease in HRCT image. The CON and GGO patterns are often appeared with the cryptogenic organizing pneumonia diseases (COPD). The GGO pattern is also often appeared in the non-specific interstitial pneumonia (NSIP). The RET pattern which sometimes includes GGO patterns is also appeared in the NSIP. The HCM pattern has more rough mesh structure rather than that of the crazy-paving, and it appeared in the idiopathic pulmonary fibrosis (IPF) or the usual interstitial pneumonia (UIP).

3.2 Texture features from Region of Interest

We introduce several texture representations proposed by Sugata et al. for features[15][1]. From the input HRCT ROI image, we calculate gray-level histogram, gray-level difference statistics, the co-occurrence matrix, run length matrix, and Fourier power spectrum, at first. After that, from these 5 quantities, we derive 39 texture statistics as the candidates for features[1]. From each of the gray-level histogram, gray-level difference statistics, Fourier power spectrum for the radial direction and for the angle, we extract contrast, variance, skewness, kurtosis, energy and entropy. From the co-occurrence matrix, we extract energy, contrast, correlation, variance, entropy. From the run length matrix, we extract short/long run emphasizes, gray level no-uniformity, run length no-uniformity, and run percentage.

3.3 Configuration of replica exchange Markov Chain Monte Carlo method

We prepare M = 7 temperature replica systems, and iterate T_{max} = 20,000 times. We use ‘libsvm’ as the linear SVM implementation. We use the default parameters for the SVM. For the CV method, we apply 10-fold CV score as the cost function H(s). The number of target class is 7, so that we adopt ‘one-versus-rest’ (OVR) classification method. The OVR method construct the class specific classifier, so that, the one classifier identify the input is belongs to the class or not.

4. Results

Fig.5 shows the result of the density of H(s) with ExMCMC method. Each figure shows the density histogram, and the horizontal axis shows the CV score, and the vertical one shows the density. The solid bar shows the histogram and the red line shows the estimated density with Gaussian kernel method. The top row shows the results for CON, GGO, HCM, RET classes, and the bottom shows the ones for the EMP, NOD, and NOR classes. The left limit of each
Fig. 5: Density of target cost function $H(s)$ for each class. The horizontal axis shows the CV error, and the vertical shows the density. In each figure, the solid bars show the density histogram, and the red curve shows the estimated density with a Gaussian kernel.

Fig. 6: Selected features for each class with top-5 CV scores. The left column shows selected features for CON, HCM, EMP, and NOR classes. The right shows for GGO, RET, and NOD classes. The black boxes show the locations of the selected features in each class.

Histogram shows the best CV score for the class. We can see the CON class is easy to classify since it has a lot of $CV = 0$ state. Moreover, the CON class is insensitive to the feature selection because it has a lot of state in the $CV = 0$ mode. On the contrary, we can see the construction of the other classifier is not so easier than the CON class. The minimum CV state has not so very few state, and we confirm the performance is very sensitive to the feature selection. Especially, we can also see the both GGO and NOD classes are hard classify since the left limit value of the histogram has the just larger than the other histograms.

Fig. 6 shows the selected features for top-5 CV scores.
The left column shows the result for CON, HCM, EMP, and NOR classes. The right one shows for GGO, RET, and NOD classes. Each horizontal axis shows the feature indices. Features from 0 to 5 come from co-occurrence matrix, from 6 to 10 come from run-length matrix, from 11 to 17 come from gray-level histograms, from 18 to 24 come from gray-difference statistics, and from 25 to 38 come from Fourier power spectrum. For CON class, the almost all features except coming from run-length matrix looks effective features. For GGO class, the gray-difference statistics, and the Fourier power spectrum in the angle direction looks important. For HCM class, gray-level histogram, gray-difference statistics, and Fourier power spectrum in the direction are important. For RET class, gray-level histograms, gray-difference statistics, and Fourier power spectrum in the radial direction are important. For EMP class, the co-occurrence matrix, and Fourier power spectrums might be important. For NOD and NOR classes, the co-occurrence matrix, and gray-level histogram might be important.

5. Conclusion & Discussion

In this research, we propose a feature selection method in the manner of Nagata’s method. The original idea of this feature selection method comes from exhaustive search, however the calculation complexity of the feature selection problem is $O(2^D)$ where $D$ means the number of total features. For small $D$, we can search the best combination with exhaustive search, but it becomes hard when $D$ becomes large. So, we introduce replica exchange MCMC method for enumeration. Applying the MCMC method, we can obtain the density curve as well as the quasi-optimal solution. We can see the how many solutions around the quasi-optimal solution, so that, we can guess the difficulties of the classification problem.

In the future work, we should compare the result with the feature selection using some classification methods with sparse prior. Introducing a sparse prior is a powerful method, however, sometimes the solutions comes from different sparse priors show different feature selection[10]. So that, we should take it carefully. In such case, the result for the ExMCMC method might be a good indicator.

Acknowledgment

We thank Professor Junji Ueno, Tokushima University. He provided several advice for this study as well as a set of high resolution HRCT image of IIPs. This work is supported by Grant-in-Aids for Scientific Research (C) 16K00328, and Innovative Areas 16H01452, MEXT, Japan.

References