Original Article

Healthc Inform Res. 2013 June;19(2):130-136.
http://dx.doi.org/10.4258/hir.2013.19.2.130
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Modified Mixture of Experts for the Diagnosis of Perfusion Magnetic Resonance Imaging Measures in Locally Rectal Cancer Patients

Sungmin Myoung, PhD
Department of Medical Information and Administration, College of Health Science, Jungwon University, Goesan, Korea

Objectives: This study demonstrates the feasibility of using a modified mixture of experts (ME) model with repeated measured tumoural K^{trans} value to perform an automatic diagnosis of responder based on perfusion magnetic resonance imaging (MRI) of rectal cancer. **Methods:** The data used in this study was obtained from 39 patients with primary rectal carcinoma who were scheduled for preoperative chemoradiotherapy. The modified ME model is a joint modeling of the ME model via the linear mixed effect model. First, we considered two local experts and a gating network, and the modified expert network as a liner mixed effect model. Afterward, the finding estimates were obtained via the expectation-maximization algorithm. All computation was performed by R-2.15.2. **Results:** We found that two experts have different patterns. The feature of expert 1 (n = 10) had a higher baseline value and a lower slope than expert 2 (n = 29). A comparison of the estimated experts and responder/non-responder groups according to T-downstaging criteria showed that expert 1 had a more effect treatment responder than expert 2. **Conclusions:** A novel feature of this study is that it is an extension of classical ME models in case of repeatedly measured data. The proposed model has the advantages of flexibility and adaptability for identifying distinct subgroups with various time patterns, and it can be applied to biomedical data which is measured repeatedly, such as time-course microarray data or cohort data. This method can assist physicians as important diagnostic decision making mechanism.

Keywords: Mixture of Experts, Classification, Magnetic Resonance Imaging, Rectal Neoplasms, Medical Decision Making

Submitted: April 19, 2013
Revised: June 14, 2013
Accepted: June 15, 2013

Corresponding Author
Sungmin Myoung, PhD
Department of Medical Information and Administration, College of Health Science, Jungwon University, 85 Munmu-ro, Goesan-eup, Goesan-gun 367-700, Korea. Tel: +82-43-830-8604, Fax: +82-43-830-8679, E-mail: smmyoung@jwu.ac.kr

I. Introduction

Recent advances in the field of biomedical informatics have led to the use of multiple models for pattern classification and regression. The incorporation of computational techniques and artificial intelligence in medicine can be applied in the classification and prediction of disease [1,2]. For example, Spilker et al. [3] used mixture model analysis to categorize benign and malignant tumor subgroups in animal models.

Among many methods, the 'divide-and-conquer' principle, which is often used to address complex problems by dividing them into simpler problems whose solution can be combined to yield a final solution, is applied for the solution of complexities in data [4].

In applying this principle, Jordan and Jacobs [5] proposed a modular neural network architecture called mixture of experts (ME). The mixture of experts contains two networks, one is called the 'experts' network, which is a population of simple linear classifiers; the other is called the 'gating' network, which mixes the outputs of the expert network. The gating network receives input vector and produces a partition of unity at each point in the input space. The outputs of expert networks are combined by a gating network simultaneously trained to stochastically select the expert that is performing the best at solving the problem [6].

To learn the algorithm for the ME model, Jordan and Jacobs [5] proposed the expectation-maximization (EM) algorithm, which is a general statistical approximation algorithm of Dempster et al. [7]. Using this algorithm, the learning process is separated to fit with the modular structure. Also, the EM algorithm can be extended to provide an effective training mechanism for the ME model based on a Gaussian probability assumption [7].

Research in the area of using the ME model has been carried for several years. Ubeyli [8] applied it to the diagnosis of breast cancer and diabetes, Subasi [9] used electroencephalography signal classification for wavelet feature extraction, Corchado et al. [1] proposed decision support in the diagnosis of leukemia patients, and Raman et al. [10] presented an infinite mixture of experts model to find an unknown number of sub-groups within a given patient cohort based on survival analysis. Myoung et al. [11] applied the ME model to find hidden subgroups of liver cirrhosis using transient elastography data in patients with chronic liver disease. Shankaracharya et al. [12] also proposed ME modeling to identify prediabetic, diabetic, and non-diabetic individuals with high accuracy in 1,415 Indian patients.

Preoperative chemoradiotherapy (CRT) has been widely used to decrease the local recurrence rate of locally advanced rectal cancer. Lim et al. [13] suggested that perfusion magnetic resonance imaging (MRI) of rectal cancer could be useful for assessing tumoural K^{trans} changes by CRT at three time points (2–5 days before CRT, end of the 2nd week of CRT, and 1–4 days before surgery). Also, they demonstrated that tumors with pre-CRT K^{trans} values tended to respond favorably to CRT, particularly in terms of downstaging criteria.

On the basis of this data, we will apply a ME model to find subgroups with respect to various time trend patterns. However, it is necessary to modify the ME framework, because the classical ME model cannot be applied to repeatedly measured data. The main concept of the modified ME model is replaced with a linear mixed model in an alternative to the expert network. Then, we apply the EM algorithm for learning the modified ME model. In this study, date obtained from 39 patients with primary rectal carcinoma who were scheduled for preoperative CRT were used to test the modified model. Thus, we expect to classify patients who have various time trend patterns.

This method has a further advantage in that a modified ME model can facilitate the classification of risk groups in medical decision making problems.

The remainder of this paper is organized as follows. In the second section, the ME architecture is briefly explained, and the perfusion MRI data is described. The EM algorithm used for the modified ME architecture is presented. Also, the modified ME model is proposed for application to repeated measures data. In the third section, the results of applying the proposed ME model to the MRI data are reported. Finally, in the last section we give some concluding remarks and discuss extensions to the proposed model.

II. Methods

1. Perfusion MRI Data Overview

Preoperative CRT has been widely used to decrease the local recurrence rates of locally advanced rectal cancer. CRT can induce complete/partial pathological response and tumour downstaging due to vascular changes and cell death [14]. Lim et al. [13] monitored the permeability change effects of CRT in rectal cancer and assessed the usefulness of perfusion MRI-based pharmacokinetic parameters as potential biomarkers to predict the response to CRT in locally advanced rectal cancer [13].

In this study, this perfusion MRI data was analyzed. The data was obtained from 39 patients who were enrolled from September 2008 to February 2010, and each patient met the following criteria: clinical MR stage T3 or T4 (the tumor could be any N stage without evidence of distant metastases on other imaging techniques).

All enrolled patients were examined using MR imaging at three time points: 2–5 days before CRT (pre-CRT MRI), at the end of the 2nd week of CRT (early-CRT MRI), and 1–4 days before surgery (post-CRT MRI). The K^{trans} (volume transfer constant) value was measured. It describes the transfer rate of the contrast agent from the blood plasma into the extravascular extracellular space (EES) by using Interactive Data Language (IDL; Research Systems, Boulder, Co, USA). Following surgery, a total of 19 patients did not show T-downstaging following CRT, while 20 patients showed T-downstaging, including 7 patients which had

Table 1. Distribution of responders and non-responders after surgery according to the downstaging standards

Downstaging standard of tumour response	Responder			Non-responder		
	cT4	cT3	Total	cT4	cT3	Total
ypT4	0	0	0	1	0	1
ypT3	7	0	7	0	18	18
ypT2	1	3	4	0	0	0
ypT1	0	2	2	0	0	0
ypT0	2	5	7	0	0	0
Total	10	10	20 (51.3%)	1	18	19 (48.7%)

cT denotes preoperative chemoradiotherapy magnetic resonance stage, ypT denotes pathological tumour stage.

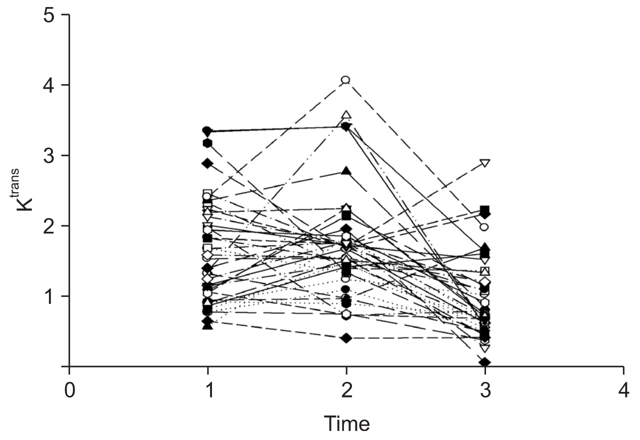

Figure 1. Temporal change in K^{trans} by preoperative chemoradiotherapy in all patients.

complete pathological remission (Table 1). Figure 1 shows the temporal changes of K^{trans} value in all patients. Lim et al. [13] concluded that perfusion MR imaging of rectal cancer can be useful for assessing tumoural K^{trans} changes as a result of CRT, and tumors with high pre-CRT K^{trans} values tend to respond favorably to CRT, particularly in terms of downstaging criteria.

2. Mixture of Expert and EM-Algorithm

In this subsection, we briefly review the ME architecture [4] and the EM algorithm [5]. Figure 2 shows a graphical representation of the ME in the expert network i = 2. This can be viewed as a tree-structured model. The ME model consists of i experts, which sit at the leaves of the tree. These experts maps the covariate space to output vector μ_i. It is assumed that different experts are solved in different local regions of the input space. Thus, this ME model uses the 'divide-and-conquer' principle, which addresses a complex tasks by dividing it into simpler subtasks.

The gating network receives the input vector x and identifies scalar outputs that are likely to approximate the partition of unity at each point in the input space. The output of the gating network is the probability g_i which weighs the contributions of each expert network. Thus, the total output of the ME model is the convex weighted sum of all the output vectors produced by the expert networks [11].

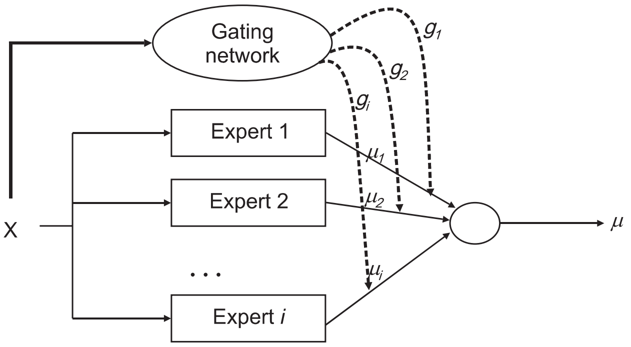

Figure 2. A mixture of experts model.

All of the expert networks are linear with a single output nonlinearity which is referred to as 'generalized linear' [5,15]. The i-th expert network produces its output μ_i as a generalized linear function of the input x:

$$\mu_i = f(W_i x), \tag{1}$$

where W_i is a weight matrix, and function $f(\cdot)$ is a fixed continuous nonlinearity. Here, $f(\cdot)$ is generally chosen to be the logistic function or the identity function. These models are smoothed piecewise analogs of the corresponding generalized linear interactive modeling (GLIM) models [5]. The gating network is also a generalized linear function, and the i-th output of the gating network g_i is the multinomial logit or softmax function of ξ_i [4]:

$$g_i = \frac{e^{\xi_i}}{\sum_{k=1}^{N} e^{\xi_k}}, \tag{2}$$

where e denotes the natural exponential, $\xi_i = v_i^T x$ denotes intermediate variables, and v_i is a weight vector. The overall output μ of the ME architecture is calculated as,

$$\mu = \sum_{k=1}^{N} g_k \mu_k \,. \tag{3}$$

the ME architecture can be given a probabilistic interpretation. The values of gating network g_i are interpreted as the multinomial probabilities associated with the decision that terminates in a regressive process that maps x to y for an input-output pair (x, y). Once the decision has been made, resulting in a choice of regressive process i, output y is then taken from a probability density $P(y|x, W_i)$, where W_i is the weight matrix of the *i-th* expert network in the model. Therefore, the total probability of observing y from x is given by mixture of the probabilities of generating y from the component densities which have multinomial probabilities:

$$P(y \mid x, \Phi) = \sum_{k=1}^{N} g_k P(y \mid x, W_k), \tag{4}$$

where Φ includes the expert network parameters and the gating network parameters. Given the total probability model, a learning algorithm for the ME architecture is considered as a maximum likelihood estimation problem. Jacobs et al. [4] as well as Jordan and Xu [16] proposed a learning algorithm for the ME architecture based on the EM framework, which is based on the idea of simplifying if only set of additional 'missing' variables.

Given an observed data set $\chi = \{(x_t, y_t)\}_{t=1}^{T}$, the EM algorithm is an iterative algorithm comprising two steps:

1) the expectation (E) step, which computes the posterior probabilities $h_i^{(t)} (i=1,\cdots,N)$, which can be interpreted as the probabilities $P(i|x_t, y_t)$ and are given for the s-th epoch as,

$$h_i^{(t)} = \frac{g(x_t, v_i^{(s)}) P(y_t \mid x_t, W_i^{(s)})}{\sum_{k=1}^{N} g(x_t, v_k^{(s)}) P(y_t \mid x_t, W_k^{(s)})}; \tag{5}$$

2) the maximization (M) step, which solves the following problems:

$$\begin{gathered} W_i^{(s+1)} = \arg\max_{W_i} \sum_{t=1}^{T} h_i^{(t)} \log P(y_t \mid x_t, W_i) \quad \text{and} \\ V^{(s+1)} = \arg\max_{V} \sum_{t=1}^{T} \sum_{k=1}^{N} h_k^{(t)} \log g_k \end{gathered} \tag{6}$$

where V is the set of all the parameters in the gating network.

Therefore, the EM algorithm yields the following learning algorithm [3]:

1) For each data pair (x_t, y_t), compute the posterior probabilities $h_i^{(t)}$ using the current values of the parameters.
2) For each expert network i, solve the maximization problem in $W_i^{(s+1)}$ with observation $\{(x_t, y_t)\}_{t=1}^{T}$ and observation weights $\{h_i^{(t)}\}_{t=1}^{T}$.
3) For the gating network, solve the maximization problem in $V^{(s+1)}$ with observation $\{(x_t, h_k^{(t)})\}_{t=1}^{T}$.
4) Iterate by using the updated parameter values.

3. Modified Mixture of Experts

As described in subsection 2, the ME model has two common approaches for selecting expert networks. For regression problems, the experts are linear. For binary classification problems, the experts are smoothed piecewise analogs of the corresponding GLIM models [5]. However, these models have constraints in that it is difficult to apply each expert network due to repeatedly measured data.

Therefore, our choice of the modified expert network is considered a linear mixed effect model for modeling purposes. For the repeated measurement problem, we consider the modified expert network as follows:

$$y_i = X_i\beta + Z_i\beta_i + \varepsilon_i, \tag{7}$$

where β is the fixed effect of time, β_i is random effect, $\varepsilon_i \sim N(0, \sigma^2)$ is the measured error. Parameter estimation in the modified ME model is a maximum likelihood learning problem, and the EM algorithm can be used to solve it.

III. Results

In this section, we report how we used the modified ME model to find subgroups with the K^{trans} value in rectal cancer. Data was used as described by Lim et al. [13]. The K^{trans} values of 39 patients who were scheduled for preoperative CRT were measured between September 2008 and February 2010. All patients were examined at three time points, and they were divided into responder/non-responder groups according to downstaging criteria. As seen in Figure 1, the overall K^{trans} value decreased slightly.

We studied a modular architecture with $k = 2$ expert networks. That is, the ME comprised two local experts and a gating network. The output of each of the i-th experts was produced via a linear mixed effect model of the input. That is, $\mu_i = f(W_i x)$ were linear mixed effect model with (x, θ_j), where θ_j is parameter vector. For the gating network, we have, which is $g_i(x, \theta_0)$ also a generalized linear function in the classical ME model, and it is used by the generic softmax function. This architecture is shown in Figure 3.

The initial values of the gating network, expert network, and variance were chosen randomly. All computations were performed by R-2.15.2 (software available at http://www.r-project.org).

Temporal changes in the K^{trans} values of all patients are shown in Table 2. The p-values were calculated by repeated measurements using ANOVA and a two-sample t-test. At the end of the 2nd week of CRT, the mean K^{trans} value of all patients was a little higher than that in the pre-CRT phase, but the difference was not statistically significant ($p = 0.4846$). In post-CRT phase, the K^{trans} value showed a significant decrease ($p < 0.0001$).

The K^{trans} values at each three time points showed a slight difference between the responders and non-responders group. The mean K^{trans} value of the responder group was higher than that of the non-responder group except for the post-CRT phase ($p = 0.6449$).

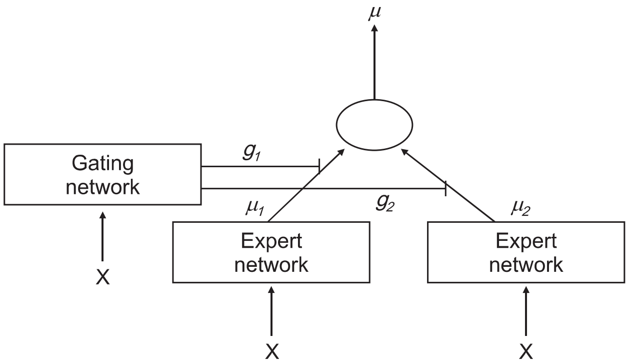

Figure 3. Configured mixture of experts structure for finding subgroups with K^{trans} value in rectal cancer.

The modified ME model with two experts was fitted to these data, resulting in the estimates $(\hat{g}_1,\hat{g}_2) = (0.2545, 0.7455)$ and $(\hat{\beta}_0,\hat{\beta}_1) = (3.4463, -0.6496)$ with expert 1, and $(\hat{\beta}_0,\hat{\beta}_1) = (1,6740, -0.2116)$ with expert 2 (Table 3). Here, $\hat{g}_1$ denotes the i-th gating network, $\hat{\beta}_0$ is the baseline value of each expert, and $\hat{\beta}_1$ is the time slope of each expert. As a result, we see that parameter of expert 1 is higher than that of expert 2. Figure 4 shows the evolution of the parameters for the first and second expert networks. The solid line represents the first parameter $\hat{\beta}_0$, and the dashed line shows the second parameter $\hat{\beta}_1$. This results shows that this modified ME model converged within 5–6 steps. Figure 5 shows the temporal changes of the K^{trans} values of the estimated two experts. In particular, the horizontal axis gives the time point (pre/early/post-CRT), and the vertical axis gives the K^{trans} value. These results along with Table 3 clearly show that expert 1 is characterized by more temporal change of the K^{trans} value than expert 2.

A comparison of the estimated experts and responder/non-responder groups according to T-downstaging criteria is shown in Table 4. In expert 1, the responder group is much larger than the non-responder group. However, the responder group is smaller than the non-responder in expert 2. This show that expert 1 is more effectively treated responders compared with expert 2.

IV. Discussion

The purpose of the present research was to apply the modi-

Table 2. Distribution of responders and non-responders after surgery according to the downstaging standards

Time point	Overall group	p-value[a]	Downstaging		p-value[b]
			Responder (n = 20)	Non-responder (n = 19)	
Pre-CRT	1.65 ± 0.76	-	1.93 ± 0.76	1.35 ± 0.67	0.0178
Early-CRT	1.74 ± 0.80	0.4846	2.08 ± 0.94	1.37 ± 0.39	0.0042
Post-CRT	1.00 ± 0.61	<0.0001	0.96 ± 0.64	1.05 ± 0.59	0.6449

Values are presented as mean ± standard deviation.

CRT: chemoradiotherapy.

[a] Calculated by repeated measures ANOVA, [b] calculated by two sample t-test.

Table 3. Parameter estimation for the modified mixture of experts architecture in K^{trans} data

Expert network	$\hat{\beta}$ ± standard error $(\hat{\beta}_0,\hat{\beta}_1)$	$\hat{g}_i$ (n, %)	$\hat{\sigma}^2$
Expert 1	(3.4463 ± 0.1869, −0.6496 ± 0.0863)	(10, 0.2545)	0.3846
Expert 2	(1.6740 ± 0.1185, −0.2116 ± 0.0523)	(29, 0.7455)	0.4196

http://dx.doi.org/10.4258/hir.2013.19.2.130

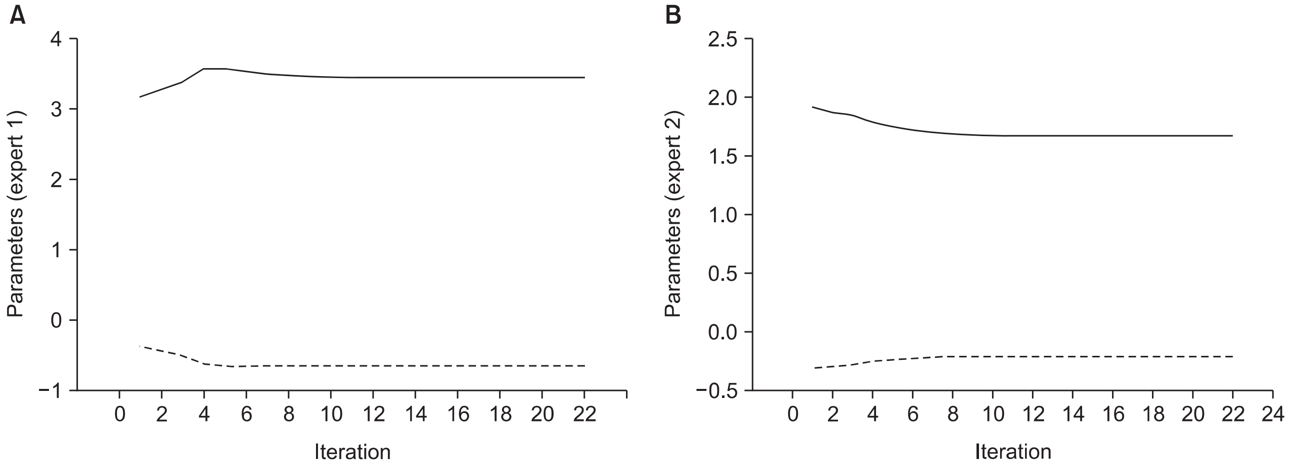

Figure 4. The performance of the modified mixture of experts model. (A) The evolution of the parameters for expert network 1 and (B) the evolution of the parameters for expert network 2.

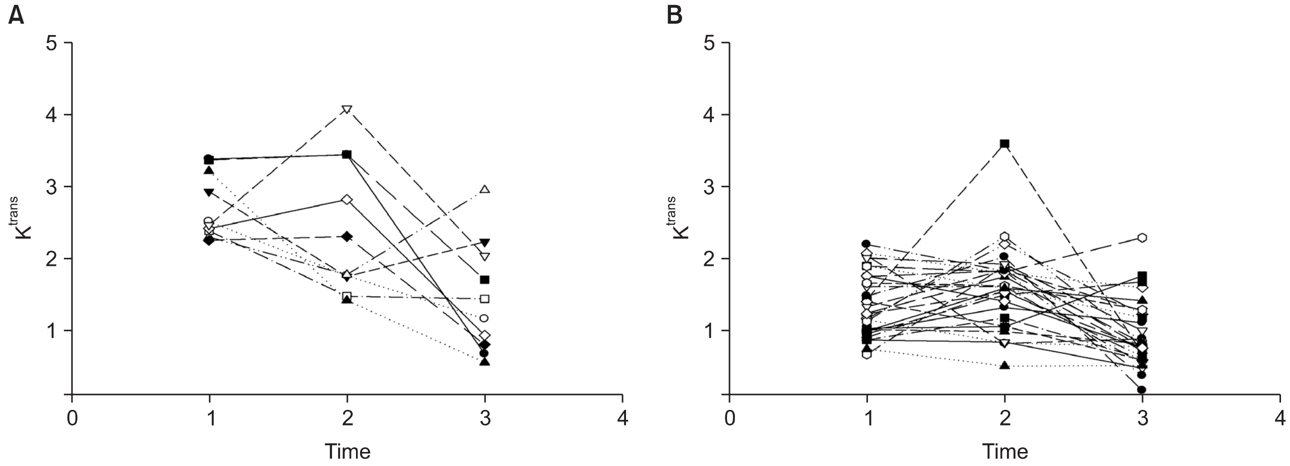

Figure 5. Temporal change in K^{trans} by preoperative chemoradiotherapy in estimated expert 1 (A) and expert 2 (B).

Table 4. Distribution of 39 patients of responder/non-responder based on T-downstaging criteria over the two estimated experts using the proposed method

Downstaging criteria	Expert 1	Expert 2
Responder	7 (70)	13 (45)
Non-responder	3 (30)	16 (55)
Total	10	29

Values are presented as number (%).

fied ME model to find subgroups with respect to various time trend patterns. In the modified ME model, linear mixed effect models were considered to take into account time trends for repeatedly measured data. The EM algorithm was used to estimate the suggested model. To classify subgroups in perfusion MRI data, two local experts and a gating network were used in the configuration of the ME architecture.

The classification results and the values of statistical parameters were used to assess the performance of the proposed ME model. We also compared the estimated experts and responder/non-responder groups according to T-downstaging criteria. In conclusion, the advantage of using the proposed ME models lies in its flexibility and adaptability for repeatedly measured data, such as K^{trans} value measured three times. The obtained results confirmed the validity of the classifiers for application in diagnostic decision support. The proposed model is expected to provide a robust method for subject classification without user intervention or bias. Future studies will be required to apply the nonlinear mixed effect model to repeatedly measured data. The number of experts was fixed to two groups (high risk/low risk) in this research, but we will be determined number of experts according to some criteria such as Akaike Information Criterion or Bayesian Information Criterion in the next study.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The author is grateful to Professor Joon Seok Lim of the Department of Radiology, Yonsei University Health System, Seoul, Korea, for providing valuable comments on this problem and data sets. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0013877).

References

1. Corchado JM, De Paz JF, Rodriguez S, Bajo J. Model of experts for decision support in the diagnosis of leukemia patients. Artif Intell Med 2009;46(3):179-200.
2. Shortliffe EH, Cimino JJ. Biomedical informatics: computer applications in health care and biomedicine. 3rd ed. New York (NY): Springer; 2006.
3. Spilker ME, Seng KY, Yao AA, Daldrup-Link HE, Shames DM, Brasch RC, et al. Mixture model approach to tumor classification based on pharmacokinetic measures of tumor permeability. J Magn Reson Imaging 2005;22(4):549-58.
4. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Comput 1991;3(1):79-87.
5. Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. Neural Comput 1994;6(2):181-214.
6. Chen K, Xu L, Chi H. Improved learning algorithms for mixture of experts in multiclass classification. Neural Netw 1999;12(9):1229-52.
7. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B Stat Methodol 1977;39(1):1-38.
8. Ubeyli ED. A mixture of experts network structure for breast cancer diagnosis. J Med Syst 2005;29(5):569-79.
9. Subasi A. EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst Appl 2007;32(4):1084-93
10. Raman S, Fuchs TJ, Wild PJ, Dahl E, Buhmann JM, Roth V. Infinite mixture-of-experts model for sparse survival regression with application to breast cancer. BMC Bioinformatics 2010;11 Suppl 8:S8.
11. Myoung S, Chang JH, Song K. A mixture of experts model for the diagnosis of liver cirrhosis by measuring the liver stiffness. Healthc Informs Res 2012;18(1):29-34.
12. Shankaracharya, Odedra D, Samanta S, Vidyarthi AS. Computational intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India. Rev Diabet Stud 2012;9(1):55-62.
13. Lim JS, Kim D, Baek SE, Myoung S, Choi J, Shin SJ, et al. Perfusion MRI for the prediction of treatment response after preoperative chemoradiotherapy in locally advanced rectal cancer. Eur Radiol 2012;22(8):1693-700.
14. Padhani AR, Liu G, Koh DM, Chenevert TL, Thoeny HC, Takahara T, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. Neoplasia 2009;11(2):102-25.
15. McCullagh P, Nelder JA. Generalized linear models. New York (NY): Chapman and Hall; 1983.
16. Jordan MI, Xu L. Convergence results for the EM approach to mixture of experts architectures. Neural Netw 1995;8(9):1409-31.

http://dx.doi.org/10.4258/hir.2013.19.2.130