PREDICTING DISEASE ACTIVITY FOR BIOLOGIC SELECTION IN RHEUMATOID ARTHRITIS

Morio YAMAUCHI[1], Kazuhisa NAKANO[2],
Yoshiya TANAKA[2] and Keiichi HORIO[1]

[1]Department of Human Intelligence Systems, Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, Japan
[2]The First Department of International Medicine, University of Occupational and Environmental Health, Kitakyushu, Japan

ABSTRACT

In this article, we implemented a regression model and conducted experiments for predicting disease activity using data from 1929 rheumatoid arthritis patients to assist in the selection of biologics for rheumatoid arthritis. On modelling, the missing variables in the data were completed by three different methods, mean value, self-organizing map and random value. Experimental results showed that the prediction error of the regression model was large regardless of the missing completion method, making it difficult to predict the prognosis of rheumatoid arthritis patients.

KEYWORDS

Rheumatoid Arthritis, Gaussian Process Regression, Self-Organizing Map

1. INTRODUCTION

Rheumatoid Arthritis (RA) causes swelling, pain and deformity of joints, and can be risks for the death in severe cases. Since the approval of infliximab that is first biologic for RA in Japan in 2003, new products have been approved one after another, making it possible to treat RA with aimed at remission [1]. However, there is no clear method or literature on biologics selection for individual cases of RA, and the establishment of such a method is an urgent issue [2]. The selection of biologics for the treatment of RA is a careful one, considering the side effects of the drugs and the patient's underlying disease. At the same time, if we have ability to predict the effect of a given biologic agent on a patient, it will obviously help in the selection of biologics.

The purpose of this study is to create a regression model for predicting the most effective biologic for RA patients from the early stages of the disease, using data accumulated over the past 15 years at the Hospital of the University Occupational and Environmental Health of Japan. The patient data to be used in this study contains many missing data, some kind of processing is required. In this article, we applied three complementary methods (mean value, self-organizing map, and random value) to the data, and compared the results with the creation of a regression model and prognosis prediction of RA patients with using the model.

David C. Wyld et al. (Eds): SPPR, NECO, GridCom, ICCSEA, SCAI, UBIC, SEMIT, MLDS - 2020
pp. 159-167, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101913

2. RELATED WORK

Kobayashi et al. evaluated the efficacy of increasing the dosage of infliximab and shortening the dosing interval of RA patients who required intensified treatment with one of the biologic agents, infliximab, and found that it increased the likelihood of inadequate response in patients with high

C-Reactive Protein (CRP) [3]. Sudo et al. statistically investigated the predictive factors of progression of joint destruction 3 years after the start of treatment in 17 RA patients on biologic agents [4]. In the literature, ARASHI status that is indicator of major joint destruction in RA and SUVmax that is the maximum radiation concentration measured from diagnostic images, were reported to be the factors most associated with the progression of joint destruction in RA at 3 years.

These studies relate specific examination items in the treatment of RA to patient outcomes, such as disease activity and joint destruction, and will undoubtedly provide useful insights into treatment strategy and follow-up. To differentiate, this study aimed to create a regression model with the items examined at the start of RA treatment and 2 weeks later as explanatory variables to directly predict the disease activity of RA patients after 6 months for each biologic agent administered.

3. RHEUMATOID ARTHRITIS DATASET

The data for RA patients used in this paper are based on the chart history of 1929 patients who had RA and were briefly admitted to the Hospital of the University Occupational and Environmental Health of Japan during the 15 years from 2003 to 2017. The Clinical Disease Activity Index (CDAI) was used to assess disease activity, with lower CDAI values indicating mild RA symptoms and higher CDAI values indicating severe disability for life and physical function.

Table 1 shows the breakdown of RA patient data by formulation and disease activity at six months. The number of variables included in the data is 55 and details are shown in Table 2.

4. METHOD

In this study, we adopted Gaussian process regression as the method of regression. SOM was used as one of the methods for interpolating the missing values. In this chapter, we describe these methods.

4.1. Gaussian Process Regression Model

Gaussian Process Regression Model is a regression model that defines input-output relationships based on Gaussian processes. One of its features is that the variance is obtained along with each prediction, which is the output of the regression model. This variance can be viewed as the confidence level of the predictions.

This confidence level of this prediction, together with the predictions output from the regression model, can be used as an aid to physicians' decision-making in selecting a biologic for RA patients.

4.2. Missing Value Imputation by SOM

SOM is one of the machine learning methods developed by Kohonen to model the visual cortex of the cortex and is classified as unsupervised learning [5]. By nonlinear mapping of data in high-dimensional space with complex correlations to low-dimensional space, we can visualize the potential features of the data. To complement the missing parts of data by self-organization map, we first collect only samples that do not have any missing parts in each variable to create a set of input signals for learning the self-organization map.

Table 1. Breakdown of RA patient data by formulation and disease activity.

	ABT	**ADA**	**CZP**	**ETN**	**GLM**	**IFX**	**TCZ**	**Tofa**
Remission	100	170	68	64	12	106	94	34
Low disease	188	166	63	92	36	86	156	37
Middle disease	99	41	28	34	15	44	72	15
High disease	22	14	7	10	5	19	28	4
SUM	409	391	166	200	68	255	350	90

Table 2. Breakdown of RA patient data by formulation and disease activity.

ITEM	**Description**
SEX	Female: 0, Male: 1
AGE	Age
STAGE	The degree of progressive joint destruction
CLASS	The degree of functional impairment
Pneumovax	Dosage of pneumococcus vaccine
Baktar	Dosage of Baktar
Foliamin	Dosage of Foliamin
Iscotin	Dosage of Iscotin
MTX	Existence of administered methotrexate
Medicine	Administered biologics
CDAI (0W*, 2W, 6M**)	Evaluation index for disease activity
SDAI (0W, 2W)	Evaluation index for disease activity
VAS (0W, 2W)	Patient's visual analogue scale
D-VAS (0W, 2W)	Physician's visual analogue scale
TJ (0W, 2W)	Number of Tender Joints
TJ28(0W, 2W)	Number of Tender Joints in 28 specified joints
SJ (0W, 2W)	Number of Swollen Joints
SJ28(0W, 2W)	Number of Swollen Joints in 28 specified joints
CRP (0W, 2W)	C-Reactive Protein value
CRP [mg/dl] (0W, 2W)	C-Reactive Protein value [mg/dl]
ESR (0W, 2W)	Blood sedimentation speed
ESR [mm/hr] (0W, 2W)	Blood sedimentation speed [mm/hr]
BAP (0W, 2W)	Osteogenic marker value
HAQ (0W, 2W)	Health Assessment Questionnaire
MS (0W, 2W)	Existence of Mitral Stenosis
GH (0W, 2W)	Global Health status value evaluated by patient
NTX (0W, 2W)	Bone metabolic marker value
BH(0W)	Body Height
BMI(0W)	Body Mass Index
BW(0W)	Body Weight
CCP(0W)	Specific Diagnostic Markers for RA
KL6(0W)	Markers of interstitial pneumonia
MMP3(0W)	Proteolytic enzymes secreted by chondrocytes
MTX(2W)	Dosage of methotrexate
PSLdose[mg/day] (0W)	Dosage of Prednisolone
QOL(0W)	Quality of Life
RF(0W)	Amount of Rheumatoid Factor

*0W: value of 0 weeks after, **6M: value of 6 months after

After learning, the relationship between each variable in the sample group used as input is discretely represented by the reference vector of neurons in the map. Then, for each sample with a missing part, we select the neuron with the most similar reference vector from the trained map, respectively. When computing the similarity, the missing parts and the corresponding components of the reference vector are ignored. The value of the component corresponding to the missing part of the reference vector of a selected neuron is an estimate of its value [6].

5. EXPERIMENT

In this section, we describe the development of regression models and experiments for predicting patient outcomes using three different datasets of RA patients with different methods of missing value imputation.

These three datasets are prepared as follows. These three datasets are normalized.

1) Data imputed by Mean Value

2) Data imputed by SOM

3) Data imputed by Random Value

A Gaussian Process Regression model was created for these datasets. The RBF kernel is specified in the kernel function as a parameter of the model.

We conducted an experiment to predict the prognostic value of disease activity CDAI in RA patients using a regression model. The CDAI, the objective variable, was used six months after the diagnosis of RA. The explanatory variables used for prediction were those at week 0 and week 2 after diagnosis.

6. RESULT AND CONSIDERATION

A plot of the predictions of the regression model for the ABT-treated patients in the dataset with missing value completion with mean value assignment is shown in Figure 1. The black plots are ideal plots with the true values of the predictions on both the vertical and horizontal axes. The red, green, blue, yellow and brown plots in the lower part of the graph are plots of CDAI-6M predictions against the test data in cross-validation, where the vertical axis is the true value and the horizontal axis is the prediction value by regression, respectively. It can be seen that most of the predictions are distributed in the region between 0 and 20 of the vertical axis.

A plot of the predictions of the regression model for the ABT-treated patient population in the dataset with missing value completion by SOM is shown in Figure 2. As with the dataset with missing completions by mean value assignment, most of the predictions are distributed in the range between 0 and 20 of the vertical axis.

A plot of the predictions of the regression model for the ABT-treated patient population in a dataset with missing value completion by random value assignment is shown in Figure 3. The results are similar to the above two plots.

These results of Figure 1, Figure 2 and Figure 3 suggest that it may be difficult to predict patient outcomes of RA patients in this study, regardless of the method used to process the missing data by regression models.

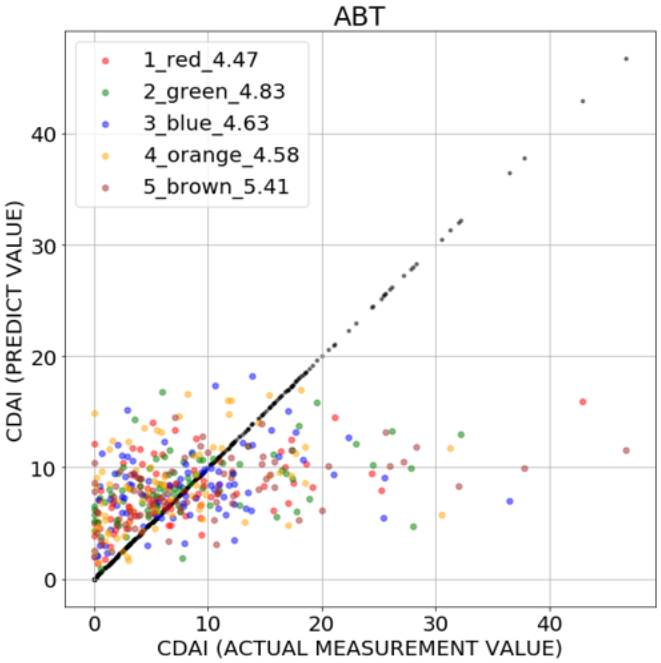

Figure 1. Results of prediction for CDAI after six months in which the missing variables are complemented by mean values.

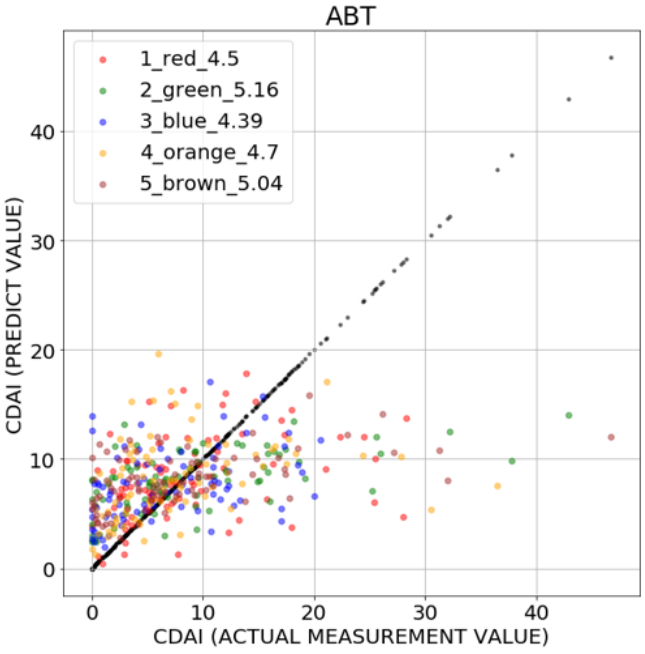

Figure 2. Results of prediction for CDAI after six months in which the missing variables are complemented by using SOM.

In support of this, a summary of the prediction error of the regression model for each missing completion method and for each administered formulation is presented in Table 3. There was no significant difference in prediction error when each formulation group was viewed by method of missing value completion.

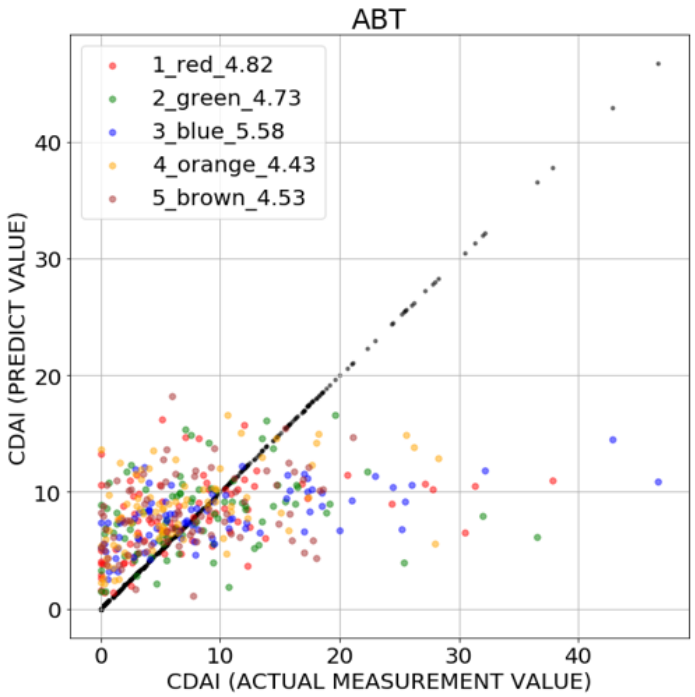

Figure 3. Results of prediction for CDAI after six months in which the missing variables are complemented by random values

.

Table 3. Prediction Error by biologics and Missing Value Imputation Method.

	Method of Missing Value Imputation		
Biologic	Mean Value	SOM	Random Value
ABT	4.76	4.76	4.82
ADA	3.63	3.66	3.64
CZP	4.31	4.26	4.43
ETN	4.09	4.33	4.26
GLM	6.95	6.19	6.12
IFX	5.89	6.03	5.95
TCZ	5.26	5.35	5.41
Tofa	4.31	4.79	4.50

These results can be attributed to the following factors.

1) Irregularities in Explanatory Variables Not Dependent on Objective Variables

The data often show large differences in the prognosis of the objective variable, disease activity, even among patients with similar explanatory variables, suggesting that regression models may not be able to capture population trends well.

There is no direct relationship between the explanatory variables at week 0 and week 2 and the disease activity at month 6, which is the objective variable, and this can be considered a factor that makes direct prediction by regression difficult, whether linear or nonlinear.

2) Insufficient adjustment of model parameters

The Gaussian process regression model used in this experiment has an advantage in that there are few parameters to be specified, but some parameters, such as the kernel function and kernel smoothness, have a significant impact on the performance. In this study, only RBF was modelled as a kernel function and kernel smoothness was modelled as an object of optimization, so further adjustments are needed

.

3) Insufficient feature engineering

In a Gaussian process regression model, the output, the objective variable, is expected to follow a Gaussian distribution. This can be achieved to some extent through data standardization and Box-Cox transformation. However, although the shape of the distribution of disease activity after six months, which was the objective variable in this experiment, approached a normal distribution, the normality assumption was not guaranteed in the Shapiro-Wilk normality test. We believe that it is necessary to consider pre-processing of data that is more suitable for the model.

4) Psycho-psychological Affected Exam Items

It is believed that pain intensity such as VAS in RA data does not directly reflect the intensity of disease activity or inflammation, but is influenced by psychological factors [7]. This may prevent regression models from capturing a direct link to disease activity, leading to prediction errors.

7. CONCLUSION

In this article, we developed a regression model and conducted experiments for predicting disease activity using data from 1929 RA patients to assist in the selection of biologics for RA. On modelling, the missing parts of the data were imputed by mean value assignment, SOM and random value assignment. Experimental results showed that the prediction error of the regression model was large regardless of the missing value imputation method, making it difficult to predict the prognosis of rheumatoid arthritis patients.

Three types of missing-value completions that were used in this study, create pseudo complete data by assigning a single value to the missing part of incomplete data. The multiple assignment method is often used in the medical field to handle incomplete data. Multiple assignment is known as a method of processing missing values that makes statistical analysis with incomplete data as statistically valid as analysis with complete data, it may be effective for the RA patient data handled in this study

.

In addition to reviewing the method for processing missing values, we will conduct interviews with physicians working in collaboration with RA in order to reduce prediction error in the model by narrowing down the variables that are important in RA treatment and those that are closely related to patient prognosis. In addition to the variables entered into the regression model as explanatory variables, the RA patient data also contain information about underlying disease and side effects of RA patients. By developing an appropriate model with that information, we aim to improve the prediction accuracy of disease activity as a patient prognosis as well as to develop a model that can predict the worsening of disease and side effects and make a decision to switch products.

REFERENCES

[1] T. Atsumi, (2017) "8. recent strategy to treat patients with rheumatoid arthritis", Nihon Naika Gakkai Zasshi, Vol. 106, No. 3, pp499-504. doi:10.2169/naika.106.499

[2] T. Takeuchi, (2020) "Present Status and Problems in Biologics for Treatment of Rheumatoid Arthritis in Japan", Nihon Naika Gakkai Zasshi, Vol. 98, No. 4, pp883-889. doi: 10.2169/naika.98.883

[3] D. Kobayashi, S. Ito, M. Unno, A. Abe, H. Otani, H. Ishikawa, …, K. Nakazono, (2017), "Effectiveness of infliximab for rheumatoid arthritis with dose escalation and shortened dosing interval", Clinical Rheumatology and Related Research, Vol.29, No. 1, pp.12-21. doi: 10.14961/cra.29.12

[4] T. Suto, Y. Yonemoto, K. Okamura, M. Tachibana, C. Okura, K. Takagishi, (2017), "Prediction of Large Joint Destruction After TNF-α Blocking Therapy in Patients with Rheumatoid Arthritis Using FDG-PET/CT and the ARASHI Scoring System",Japanese Journal of Joint Diseases, Vol.36, No. 4, pp.467-473. doi:10.11551/jsjd.36.467

[5] T. Kohonen, (1982) "Self-organized formation of topologically correct feature maps", Biological Cybernetics, Vol. 43, pp.59-69

[6] Y. Kikuchi, N. Okada, Y. Tsuji and K. Kiguchi, (2013) "An Estimating Method for Missing Data by Using Multiple Self-Organizing Maps", Transactions of The Japan Society of MechanicalEngineering, Vol.79, No.806, pp3465-3473. doi:10.1299/kikaic.79.3465

[7] N. Shimahara, H. Uchiyama, Y. Jouko, K. Akamatsu, N. Sawada, Y. Tanaka and S. Nakao, (2018) "Relationship between pain symptoms, functional impairment, and psychophysiological problems of rheumatoid arthritis patients using biologic drugs: importance of psychosocial evaluation", Clinical Rheumatology and Related Research, Vol. 30, No. 3, pp154-165. doi: 10.14961/cra.30.154