International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9858-9862
© Research India Publications. http://www.ripublication.com

A Study on Prediction of Rheumatoid Arthritis Using Machine Learning

Jihyung Yoo *, Mi Kyoung Lim *, Chunhwa Ihm, Eun Soo Choi***, Min Soo Kang*****
**Dept. of Internal Medicine, Eulji University Hospital, seo-gu dunsan-dong Daejeon 35233, Korea.*
***Dept. of Laboratory Medicine, Eulji University Hospital, seo-gu dunsan-dong Daejeon 35233, Korea.*
**** Dept. of Medical IT , Eulji University, Gyunggi-do 13135, Korea.*

Co-corresponding author: Chunhwa Ihm, MD/Ph.D.

Co-corresponding author: Min Soo Kang, Ph.D.

Abstract

Recently, the aging society in our society is one of the most serious problems in health and medical care. In an aging society, rheumatic disease is more common than any other disease, and rheumatism is a pain in the musculoskeletal system that lowers the quality of life of patients. It is very important to predict patients who will develop rheumatic diseases in terms of quality of life. In this study, clinical data were analyzed to predict patients with rheumatoid arthritis. Clinical data were analyzed for RA Factor, Anti CCP, SJC, and ESR factors to determine rheumatic disease. Data analysis was performed using the k-means algorithm to periodically study the threshold values of the rheumatoid factor, anti-CCP, SJC, and ESR factors, and predicted that if either RF or AC were positive, rheumatoid disease could occur. In this paper, we use machine learning to predict rheumatic diseases by using four factors for diagnosis of rheumatic diseases and it will help to predict rheumatic diseases using artificial intelligence in the future.

Keywords: Machine Learning, K-means Cluster, Rheumatic Arthritis, Unsupervised learning

INTRODUCTION

In Korea, although the social and economic environment has improved much more than in the past, the frequency of rheumatic diseases is increasing due to aging society. With the entry of the aging society, diseases caused by aging are closely related to the deterioration of quality of life, leading to an increase in social and economic costs. The concept of health also emphasizes the quality of life, not the disease. The World Health Organization defines the quality of life as a degree of acceptance in the culture and value system in which individuals live. Especially, osteoarthritis and rheumatoid arthritis caused by aging occupy the second place in diagnosis of chronic diseases. Although the exact cause of rheumatoid arthritis has not been elucidated yet, genetic and environmental factors interact in a complex way. It is a chronic inflammatory disease that affects one or more joints in the joint [1]. It causes pain, swelling, joint deformity, dysfunction, and autoimmune phenomenon is known as a major mechanism. In addition, symptoms other than joints are diseases that can invade whole body including anemia, dry syndrome, subcutaneous nodule. In the past, the goal of treating rheumatoid arthritis was to improve symptom. However, new drugs based on the morphology of disease have been developed and the importance of early diagnosis and early treatment has been emphasized. Therefore, early detection and treatment of rheumatoid arthritis can increase the chance of cure. There are several factors to diagnose rheumatoid arthritis. Among them, the values of Rheumatoid Factor, Anti CCP, SJC, and ESR are typical for judging rheumatoid arthritis. The k-means algorithm was used to predict the disease with four factors.

Machine learning has become a core technology for data mining and modeling for medical research as well as for IT. Already, machine learning has been used around the world, including Google's Alpha, Tensorflow, Microsoft Azure and IBM's Watson In this paper, we used clustering, a representative method of Unsupervised learning, to predict rheumatic disease patients in advance.

SYMPTOMS OF RHEUMATIC DISEASES

Rheumatoid arthritis is associated with pain in several joints and symptoms such as stiffness gradually appear over several weeks. More than 2/3 of patients experience fatigue, anorexia, ambiguous muscle and joint symptoms. These symptoms are difficult to diagnose because of rheumatic diseases. Rheumatoid arthritis is characterized by inflammation of the hands and wrists and morning stiffness. This means that the skill, experience and characteristics of the costume will affect the accuracy of the diagnosis. The doctor's diagnosis is diagnosed by physical examination, laboratory findings, and imaging findings. Rheumatoid arthritis is diagnosed by several steps and procedures. If you came to the stage of diagnosis, people would have already suffered. It is very important to predict rheumatic diseases in advance in order to raise the quality of life of the elderly in the age of aging. In this study, we studied the support of rheumatic disease diagnosis based on machine learning. The standard diagnostic criteria of ACR (American College of Rheumatology) established in 1987 as a standard for the diagnosis of rheumatic diseases are widely used and the diagnostic criteria are shown in Table 1.

Table 1. The 1987 revised criteria for the classification of rheumatoid arthritis

Criterion	Definition
1. Morning stiffness	Morning stiffness in and around the joints, lasting at least 1 hour before maximal improvement
2. Arthritis of 3 or more joint areas	At least 3 joint areas simultaneously have had soft tissue swelling or fluid (not bony overgrowth alone) observed by a physician. The 14 possible areas are right or left PIP, MCP, wrist, elbow, knee, ankle, and MTP joints
3. Arthritis of hand joints	At least 1 area swollen (as defined above) in a wrist, MCP, or PIP joint
4. Symmetric arthritis	Simultaneous involvement of the same joint areas (as defined in 2) on both sides of the body (bilateral involvement of PIPs, MCPs, or MTPs is acceptable without absolute symmetry)
5. Rheumatoid nodules	Subcutaneous nodules, over bony prominences, or extensor surfaces, or in juxta articular regions, observed by a physician
6. Serum rheumatoid factor	Demonstration of abnormal amounts of serum rheumatoid factor by any method for which the result has been positive in < 5% of normal control subjects
7. Radiographic changes	Radiographic change typical of rheumatoid arthritis on posteroanterior hand and wrist radiographs, which must include erosions or unequivocal bony decalcification localized in or most marked adjacent to the involved joints (osteoarthritis changes alone do not qualify)

However, this diagnostic method in Table 1 is not suitable for diagnosing early-onset rheumatoid arthritis patients, and the criteria for diagnosing rheumatoid arthritis patients by ACR / EULAR (European League Against Rheumatism) was revised in 2010 [2, 3]. Thus, with the help of a physician of the rheumatology physician, clinical data of Rheumatoid factor, Anti CCP, SJC and ESR were used to make early prediction of rheumatic patients. Thus, with the help of a physician of the rheumatology physician, clinical data of Rheumatoid factor, Anti CCP, SJC and ESR were used to make early prediction of rheumatic patients. Ziad Obermeyer, M.D., M.Phil of Harvard Medical School, and Ezekiel J. Emanuel, MD, Ph.D. of the University of Pennsylvania have stated that in the future, big data, machine learning technology can greatly predict the diagnosis of doctors [4].

CLUSTERING

Cluster analysis is one of the unsupervised learning analysis techniques, in which the data is given to the computer without prior knowledge and the command analysis method is "bundle similar things together". Without a specific hypothesis, there is no target variable with a hidden pattern. There are two representative clustering techniques. Non-hierarchical classification of N constituent factors into M clusters. The K-means algorithm is a typical method in which given data is grouped into k clusters, minimizing the variance of the distance difference with each cluster. Hierarchical nested clusters create several nested clusters in the process of clustering. Clustering is a fuzzy kind of clustering algorithm exists. The application of clustering is not only diverse, but it can also be applied to various algorithms in most cases. In cluster analysis, we classify four factors into four variables [5, 6].

We use k-means clustering analysis, which is a representative method of partitioning cluster analysis. K-Means clustering is a typical unsupervised learning of machine learning. For K-Means clustering, a centroid is selected and the classification is performed using the distance between the centroid and adjacent data [7]. The next step is to repeat the process of setting and sorting centroids, which are more centrally located, and end the work if no further classification is possible. K-means clustering can be restored by minimizing the Euclidean distance from the center point and by performing the initial erroneous merging algorithm repeatedly [8]. In addition, there is a feature that a meaningful data structure can be found without prior knowledge of a given data.

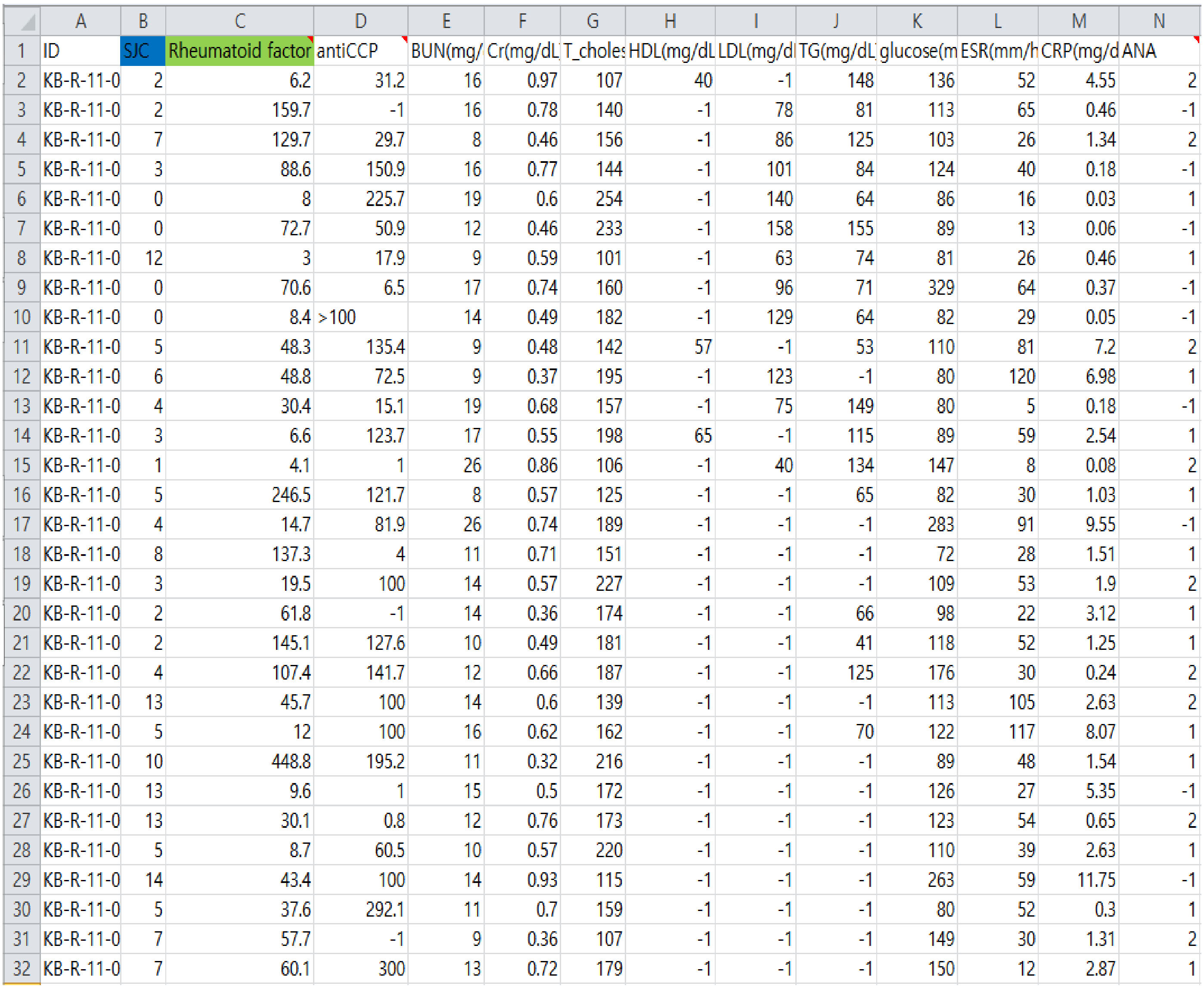
SIMULATION

Subject of study

Korea is aging at the fastest pace in the history of the world and expects the expected number of super-aged population to increase by 24.3% in 2030 and 40% in 2060 by year. Rapid population aging is emerging as a pending issue not only of individuals' physical and mental health such as the elderly, but also of changes in society as a whole, such as health care, politics and economy, and new crises [9]. The increase in the elderly population due to rapid aging leads to the financial burden of medical expenses and health insurance and leads to the deterioration of quality of life. In particular, rheumatoid arthritis causes serious problems with pain and inflammation [10]. In order to predict patients with rheumatic disease, we conducted a study based on clinical data from 60 anonymous rheumatic patients provided by the eulji university hospital. To diagnose a patient with rheumatoid arthritis, a diagnosis is made based on symptom, laboratory result, imaging finding. However, in this study, four factors among many data were diagnosed. Patients with rheumatoid arthritis were assessed for rhumatoid factor> 7, anti CCP> 18, SJC> 4, and ESR> 25. Figure 1 shows patient data including each factor.

Method of study

In order to predict rheumatic diseases, 60 anonymous data were randomly clustered into four factors. We chose the k-means algorithm to verify how reliable the results of future studies are in patients with rheumatic patients, although the amount of data is small. Four of the 60 parameters were randomly clustered. Cluster 4 was selected. The reason for selecting four clusters is that they are comparative analysis results using four factors. To calculate the centroid of the cluster, we first set it randomly and set it to the nearest centroid by adding the number of times. Figure 2 shows the visualization of K-means model based on R program.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ID	SJC	Rheumatoid factor	antiCCP	BUN(mg/	Cr(mg/dL	T_choles	HDL(mg/dL	LDL(mg/dl	TG(mg/dL	glucose(m	ESR(mm/h	CRP(mg/d	ANA
2	KB-R-11-0	2	6.2	31.2	16	0.97	107	40	-1	148	136	52	4.55	2
3	KB-R-11-0	2	159.7	-1	16	0.78	140	-1	78	81	113	65	0.46	-1
4	KB-R-11-0	7	129.7	29.7	8	0.46	156	-1	86	125	103	26	1.34	2
5	KB-R-11-0	3	88.6	150.9	16	0.77	144	-1	101	84	124	40	0.18	-1
6	KB-R-11-0	0	8	225.7	19	0.6	254	-1	140	64	86	16	0.03	1
7	KB-R-11-0	0	72.7	50.9	12	0.46	233	-1	158	155	89	13	0.06	-1
8	KB-R-11-0	12	3	17.9	9	0.59	101	-1	63	74	81	26	0.46	1
9	KB-R-11-0	0	70.6	6.5	17	0.74	160	-1	96	71	329	64	0.37	-1
10	KB-R-11-0	0	8.4	>100	14	0.49	182	-1	129	64	82	29	0.05	-1
11	KB-R-11-0	5	48.3	135.4	9	0.48	142	57	-1	53	110	81	7.2	2
12	KB-R-11-0	6	48.8	72.5	9	0.37	195	-1	123	-1	80	120	6.98	1
13	KB-R-11-0	4	30.4	15.1	19	0.68	157	-1	75	149	80	5	0.18	-1
14	KB-R-11-0	3	6.6	123.7	17	0.55	198	65	-1	115	89	59	2.54	1
15	KB-R-11-0	1	4.1	1	26	0.86	106	-1	40	134	147	8	0.08	2
16	KB-R-11-0	5	246.5	121.7	8	0.57	125	-1	-1	65	82	30	1.03	1
17	KB-R-11-0	4	14.7	81.9	26	0.74	189	-1	-1	-1	283	91	9.55	-1
18	KB-R-11-0	8	137.3	4	11	0.71	151	-1	-1	-1	72	28	1.51	1
19	KB-R-11-0	3	19.5	100	14	0.57	227	-1	-1	-1	109	53	1.9	2
20	KB-R-11-0	2	61.8	-1	14	0.36	174	-1	-1	66	98	22	3.12	1
21	KB-R-11-0	2	145.1	127.6	10	0.49	181	-1	-1	41	118	52	1.25	1
22	KB-R-11-0	4	107.4	141.7	12	0.66	187	-1	-1	125	176	30	0.24	2
23	KB-R-11-0	13	45.7	100	14	0.6	139	-1	-1	-1	113	105	2.63	2
24	KB-R-11-0	5	12	100	16	0.62	162	-1	-1	70	122	117	8.07	1
25	KB-R-11-0	10	448.8	195.2	11	0.32	216	-1	-1	-1	89	48	1.54	1
26	KB-R-11-0	13	9.6	1	15	0.5	172	-1	-1	-1	126	27	5.35	-1
27	KB-R-11-0	13	30.1	0.8	12	0.76	173	-1	-1	-1	123	54	0.65	2
28	KB-R-11-0	5	8.7	60.5	10	0.57	220	-1	-1	-1	110	39	2.63	1
29	KB-R-11-0	14	43.4	100	14	0.93	115	-1	-1	-1	263	59	11.75	-1
30	KB-R-11-0	5	37.6	292.1	11	0.7	159	-1	-1	-1	80	52	0.3	1
31	KB-R-11-0	7	57.7	-1	9	0.36	107	-1	-1	-1	149	30	1.31	2
32	KB-R-11-0	7	60.1	300	13	0.72	179	-1	-1	-1	150	12	2.87	1

Figure 1. Example of data set for predictive

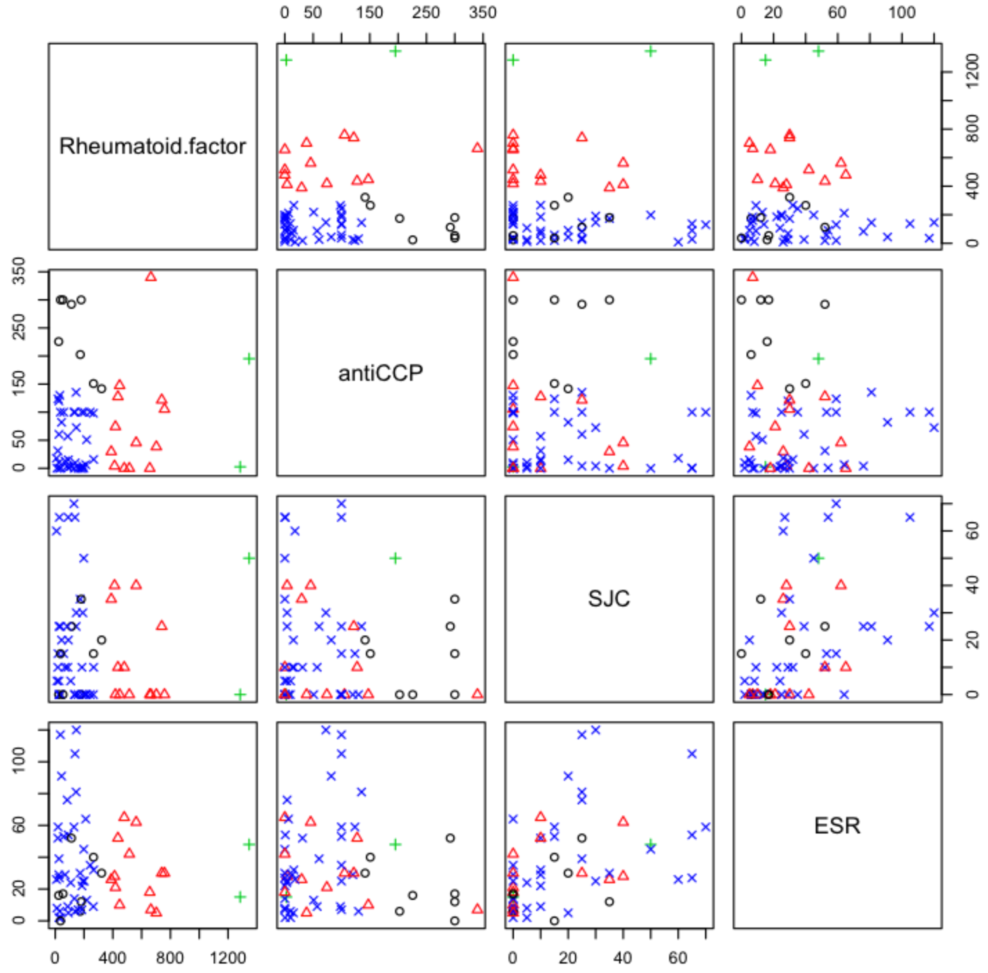
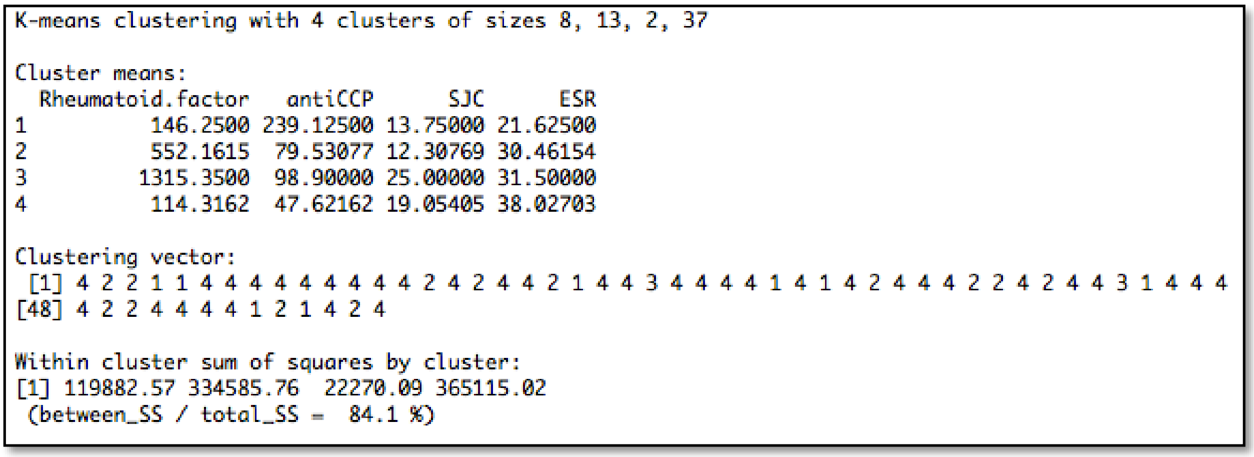


Figure 2. Visualization of K-means model

Simulation results

This study was based on data from patients with rheumatoid arthritis. Four factors were used to predict the relationship between the two factors in order to predict the rheumatic patient. One factor may be rheumatoid disease, but the results of the clustering showed that at least two cases were rheumatic patients. Figure 3 shows the Description Model of the object variable that contains the clustering result.



```
K-means clustering with 4 clusters of sizes 8, 13, 2, 37

Cluster means:
  Rheumatoid.factor   antiCCP      SJC      ESR
1          146.2500 239.12500 13.75000 21.62500
2          552.1615  79.53077 12.30769 30.46154
3         1315.3500  98.90000 25.00000 31.50000
4          114.3162  47.62162 19.05405 38.02703

Clustering vector:
 [1] 4 2 2 1 1 4 4 4 4 4 4 4 4 4 2 4 2 4 4 2 1 4 4 3 4 4 4 4 1 4 1 4 2 4 4 4 2 2 4 2 4 4 3 1 4 4 4
[48] 4 2 2 4 4 4 4 1 2 1 4 2 4

Within cluster sum of squares by cluster:
[1] 119882.57 334585.76  22270.09 365115.02
 (between_SS / total_SS =  84.1 %)
```

Figure 3. Description Model for k-means clustering results

Figure 3 shows the factor values, average values, and cluster estimations for each cluster. As a result, the result was 84.1%

CONCLUSION

In this study, we conducted a study to predict patients with rheumatoid arthritis. To predict patients with rheumatism, k = 4. Clarification evaluation results of 84% or more were obtained through the explanatory model. It is found that the result of selecting the two factors and finding the correlation is higher than the case of selecting only one parameter. The K-Means algorithm showed that rheumatoid arthritis can be predicted as two of four factors. In an aging society, rheumatoid disease is a disease that severely affects patients' quality of life. However, if we can predict the rheumatoid arthritis based on the results of this study, we can improve the quality of life.

ACKNOWLEDGMENT

“This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2017R1D1A1B03034411)”

REFERENCES

[1] Henk Visser, Saskia le Cessie, Koen Vos, Ferdinand C. Breedveld, and Johanna M. W. Hazes, How to Diagnose Rheumatoid Arthritis Early, ARTHRITIS & RHEUMATISM,Vol. 46, No. 2 (2002), 357–365.

[2] Sung-Hwan Park, New diagnostic method of rheumatoid arthritis, The Korean Journal of Medicine, Vol. 76, No.1 (2009)7-11.

[3] Jung-Soo Song, New Classification Criteria for Rheumatoid Arthritis, The Korean Journal of Medicine: Vol. 87, No. 4 (2014), 383-388.

[4] Ziad Obermeyer, Ezekiel J. Emanuel, Predicting the Future — Big Data, Machine Learning, and Clinical Medicine, The New England Journal of Medicine, Vol.375, No.13 (2016), 1216-1219.

[5] Yong Gyu Jung, Min Soo Kang, Jun Heo, Clustering performance comparison using K-means and expectation maximization algorithms, Biotechnology & Biotechnological Equipment, Vol. 28, No. S1 (2014), 45-48.

[6] Min-Soo Kang, Yong-Gyu Jung, Du-Hwan Jang, A Study on the Search of Optimal Aquaculture farm condition based on Machine Learning, The Journal of The Institute of Internet, Broadcasting and Communication (IIBC) Vol. 17, No. 2 (2017), 135-140.

[7] Jae-Gyun Park, Eun-Soo Choi, Min-Soo Kang* , Yong-Gyu Jung, Dropout Genetic Algorithm Analysis for Deep Learning Generalization Error Minimization, International Journal of Advanced Culture Technology Vol.5 No.2 (2017), 74-81.

[8] Beom-Joo Park, Min-Soo Kang, Minho Lee, Yong Gyu Jung, A Study on Efficient Memory Management Using Machine Learning Algorithm, International Journal of Advanced Smart Convergence Vol.6 No.1 (2017), 39-43

[9] Chen Lin, Elizabeth W. Karlson, Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records, PLOS, Vol.8, Issue.8 (2013), 1-10.

[10] Goo Joo Lee, Byung-Mo Oh, Keewon Kim, Sang Yoon Lee, Sewoong Chun, Tai Ryoon Han, K-means Cluster Analysis on Care Status of Injured Workers with Stroke According to Discharge Disposition Patterns, Brain & Neuro Rehabilitation Vol. 4, No. 2, (2011), 132-136.