Developing a Hybrid Decision Support Model to Discover Evidence Based Knowledge of the Elderly with Depression

Myonghwa Park[1], Chang Sik Son[2] and Sun Kyung Kim[1]

[1] *College of Nursing, Chungnam National University, Daejeon, Republic of Korea*
[2] Biomedical Informatics Technology Center, *Keimyung University, Daegu, Republic of Korea*
mhpark@cnu.ac.kr, csson@kmu.ac.kr, rlatjsrud03@naver.com

Abstract

Data mining is the process to extract hidden patterns from enormous amount of data that is commonly used in a range of areas including marketing, fraud detection, scientific discovery as well as health care. The study was conducted to ensure high accuracy in assessing of elderly depression and to build useful decision rules by developing a very reliable evidence based decision support model with the combination of statistical analysis and decision tree algorithms. A large data set of 2008 Korean Elderly Survey (KES) was used consisted of 14,970 elderly data. Having depression as target variable, input variables were demographic, health related and socioeconomic characteristics of the Korean elderly population. Statistical analysis was conducted as a feature selection procession that includes the Chi-square, Fisher's exact test, the Mann-Whitney U-test and Wald logistic regression Using the C5.0 decision tree algorithm of Clementine 12.0, the final decision support models were built and C5.0 tree showed a high accuracy level of 81.6%. The decision model developed in this study can improve healthcare providers' ability in making decisions, increasing vigilance with suspected depression in elderly population.

Keywords: *Data mining, Logistic regression, Depression, Aged*

1. Introduction

Data generated by healthcare settings are not only complex but also voluminous to be processed and analyzed by traditional statistical methods [1]. Data mining is the process to extract hidden patterns from enormous amount of data that is commonly used in a range of areas including marketing, fraud detection, scientific discovery as well as health care [2]. Therefore, it can be an effective solution to improve decision-making by discovering patterns and trends in large amounts of complex healthcare data [3]. Understandings gained from data mining can influence cost and managing efficiency while maintaining a high level of care [4].

Depression has been known to be a major contributor to healthcare costs in the elderly populations, causing overwhelming burden on the healthcare system [5]. It is reported between 20% and as high as 50% of elderly population in Korea is suffering from depression and the prevalence of disease is predicted to increase further due to the population aging [6]. According to the findings of previous studies, old adults with depressive symptom, with or without the appearance of depressive disorder, showed poorer functioning than that of other people with chronic medical conditions such as heart and lung disease, arthritis, hypertension and diabetes only [6]. The perception of poor health among depression patients makes influence on health status that patient with poor self-rated health showed poor functioning. Presence of chronic disease is also often viewed as a risk factor that depression was more common in those. There are striking effects of depression on consumption of medical care. The search for physical explanation follows normally, causing unnecessary increase in medical utilization rates. Furthermore, co-occurred depression with other medical conditions worsens the disease condition of patients and their adherence to treatments, lessening chances for improvement or recovery [7].

Old adults experience many losses in their life which include loss of health, loss of social role, financial power and relationship with their loved ones. Having experienced all those losses, elderly became more vulnerable to depressions. A number of factors are involved in the elderly depression not only socio-demographic factors such as age, marriage status, education level, socio-economic status, occupation, living status and religion, but also functional level and social support are associated with depression. The importance of the early diagnosis and proper treatment of elderly depression can be summarized as first, depression is treatable, depression can be risk factor of dementia and symptom of dementia and depression should be distinguished [8].

Using large database, application of data mining can result in the diagnosis and prognosis and even the discovery of hidden patterns in depression [9]. It is therefore required to build decision tree model of depression for the better understanding of characteristics of elderly depression and prediction of its risk factors.

1.1. Purpose of Study

The aim of this study is to develop an evidence based decision support model with the combination of statistical analysis and decision tree algorithms to ensure high accuracy in screening of elderly with depressive disorder and to build useful decision rules. The specific purposes of the study follow as;

1. To identify the variables related to depression in elderly
2. To identify the best modeling method among data mining tools, testing the discriminatory power of decision support models.
3. To develop tree structured model to describe the characteristics of the elderly with depression

2. Method

2.1. Data Source

This study used the data from the 2008 national study of Korean elderly to explore the characteristics of the elderly with depression in Korea with the Institutional Review Board (IRB) approval. The 2008 national study of Korean elderly consisted of data from 15,146 older adults (Male: 6,452, Female: 8,694) over 60 years who lived in community and investigated their health and life style condition, and welfare status [10].

2.2. Decision Tree Models and Statistical Data Analysis

Figure 1 shows the scheme of the decision support modeling that was based on statistical analysis (multivariate analysis) and the Wald forward logistic regression with entry and removal criteria of 0.05 and 0.10 each for the identification of characteristics of elderly depression.

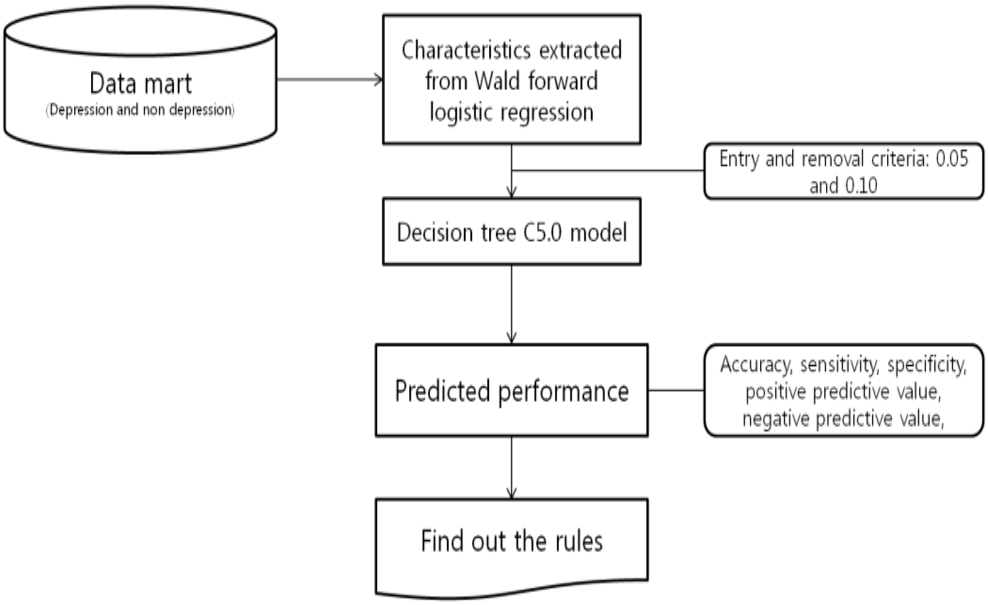
Data mart
(Depression and non depression)
Characteristics extracted from Wald forward logistic regression
Entry and removal criteria: 0.05 and 0.10
Decision tree C5.0 model
Predicted performance
Accuracy, sensitivity, specificity, positive predictive value, negative predictive value,
Find out the rules

Figure 1. Scheme of the Decision Support Modeling

The target variable was depression and 58 variables were used as input variables. Feature selection node was used based on the p-value for categorical prediction were applied to get tree structured models. Analysis node was used to compare the percentage of correct and wrong classification of the models. ROC (Receiver Operating Characteristcs) curves were drived to see the predictive performance of the models. The performance of the models was evaluated and C5.0 Tree was selected to design the final decision support models. The C5.0 Tree node builds either a decision tree or a rule set. This approach provides a very simple representation of accumulated knowledge, facilitating a clinical decision making process. The model selects the best decision node that separates the different classes from the empirical data. It works by first splitting the sample based on the field that provides the maximum information gain. Each subsample is defined by the first split then those subsamples split again based on a field difference. The whole process repeats until the subsamples are not able to split any further. At the final stage, the splits in the lowest level are reexamined, filtering those that do not have significant contribution to the value of the model are removed or pruned.

The decision tree model used in this study was built with C5.0 Tree component using the default experimental parameters of Clementine version 12.0 (SPSS Inc., Chicago, IL, USA). The rules were evaluated with the testing data set for their predictability. Statistical analysis was performed using SPSS 18.0 for Windows (SPSS Inc., Chicago, IL, USA). Univariate correlations were evaluated using the Chi-square test or Fisher’s exact test, which are appropriate for categorical data [12]. Kolmogorov-Smirnov test were conducted first to test normality and the variables were analyzed further by either of Student t-test and Mann-Whitney U-test. A two tailed $p<0.05$ was selected as the level of statistical significance. In the multivariate analysis, Wald forward selection was used with entry and removal criteria of 0.05 and 0.10. The results of modeling were expressed as the odd ratios (OR) with 95% confidence intervals.

3. Results

3.1. Demographic Characteristics of the Target Population

Demographic characteristics of participants are shown in the Table 1. Using depression screening tool, participants scored between 8 and 15 were screened as depression. Of the

14,970 subjects in the data set, 4,423(29.54%) had depression, while 10, 547 (70.46%) had no sign of depression. Significant differences were observed in terms of gender, age, education level, marital status, living arrangement and work status. When compared to non-depression group, there were higher proportion of female population and about half of elderly with depression had the lowest education level, between 0 and 5 years. Elderly with depression were older, were more likely to live alone and were less likely to have a job.

Table 1. Demographic Characteristics of the Target Population

Variables	Depression score (n=14970)		*p*
	No depression (0-7) (n=10547)	Depression (8-15) (n=4423)	
Gender (%)			0.000
Male	4663 (44.2)	1449 (32.8)	
Female	5884 (55.8)	2974 (67.2)	
Age	70.03±6.75	72.80±7.22	0.000
Education year (%)			0.000
0-5	2663 (25.2)	2027 (45.8)	
6-9	4254 (40.3)	1588 (35.9)	
10-12	1540 (14.6)	423 (9.6)	
13-15	1373 (13.0)	275 (6.2)	
>15	717 (6.8)	110 (2.5)	
Marital status (%)			0.000
Never married	23 (0.2)	30 (0.7)	
Divorce, Separation by death	7157 (67.9)	2107 (47.6)	
Married	3367 (31.9)	2286 (51.7)	
Living arrangement (%)			0.000
Living alone	1986 (18.8)	1477 (33.4)	
With spouse	5101 (48.4)	1469 (33.2)	
With adult children	2986 (28.3)	1214 (27.4)	
Work status (%)			0.000
Unemployed	6299 (59.7)	3504 (79.2)	
Employed	4248 (40.3)	919 (20.8)	

3.2. Performance of Models based on Logistic Regression

The performance of the models was evaluated using six standard measures including accuracy (ACC), sensitivity (SENS), positive predictive value (PPV), negative predictive value (NPV), and the area under the ROC curve (AUC) (Table 2). The AUC of the models was 75.3% and ACC, SENS, SPEC, PPV, and NPV, were 81.6%, 60.0%, 90.6%, 73.3%, and 84.4%, respectively.

Table 2. Performance of Classification by Decision Tree Model

Methods	ACC	SENS	SPEC	PPV	NPV	AUC	no. rules
Logistic regression Based Decision Tree Model	81.6	60.0	90.6	73.3	84.4	75.3	11

ACC, accuracy; SENS, sensitivity; SPEC, specificity; PPV, positive predictive value; NPV, negative predictive value; AUC, area under ROC curve

3.3. Analysis of Decision Support Model based on Multivariate Analysis

After feature selection, 23 variables were used as final input variables. Ten out of 23 variables, including overall satisfaction of living, average life satisfaction score, limitation in ADL and IADL, nutritional status score, exercise capacity score, perceived health status and perceived financial status were selected using the C5.0 tree algorithm. Cut off points of each item were determined by the decision tree algorithm that the criteria for dichotomizing the continuous variables were all statistically significant (<0.05). The results were summarized in Table 3 and the decision support model is shown in Figure 2.

Table 3. Cut-off Points Determined by the Decision Algorithm

Levels	OR (95% CI)	*p*
Level 0 (Root node)		
Overall life satisfaction (moderate; high; very high) or (low; very low)	10.728 (9.794–11.752)	0.000
Level 1		
Average life satisfaction score ≤ 3.462	0.161 (0.145–0.178)	0.000
Average life satisfaction score > 3.462	6.225 (5.615–6.901)	0.000
Perceived health status (bad; very bad) or (moderate; good; very good)	5.377 (4.393–6.580)	0.000
Level 2		
Limitation in ADL (no or yes)	4.194 (3.406–5.164)	0.000
Nutritional status score ≤ 4	0.292 (0.241–0.355)	0.000
Nutritional status score > 4	3.424 (2.820–4.158)	0.000
Average life satisfaction score ≤ 2.944 or > 2.944	4.241 (2.776–6.479)	0.000
Level 3		
Nutritional status score ≤ 5 or > 5	3.211 (2.794–3.691)	0.000
Overall life satisfaction (low)	0.177 (0.106–0.295)	0.000
Level 4		
Average life satisfaction score ≤ 3.214	0.336 (0.283–0.399)	0.000
Average life satisfaction score > 3.214	2.975 (2.507–3.531)	0.000
Exercise capacity score ≤ 73.333 or > 73.333	3.230 (2.422–4.308)	0.000
Limitation in IADL	3.310 (2.408–4.549)	0.000
Level 5		
Limitation in IADL (no or yes)	2.130 (1.599–2.838)	0.000
Nutritional status score ≤ 10	0.321 (0.207–0.498)	0.000
Nutritional status score > 10	3.119 (2.010–4.841)	0.000
Days of having sound sleep per week (7 day; 3-4 days; 5-6 days)or (1-2 days)	2.570 (1.732–3.815)	0.000
Level 6		
Health status (bad; very bad) or (good; moderate) or (very good)	3.618 (1.670–7.840)	0.000
Limitation in IADL	2.200 (1.619–2.988)	0.000
Level 7		
Perceived financial status (rich; poor) or (moderate)	2.313 (1.278–4.188)	0.006
Perceived financial status (rich; poor; moderate) or (poor)	2.205 (1.440–3.375)	0.000

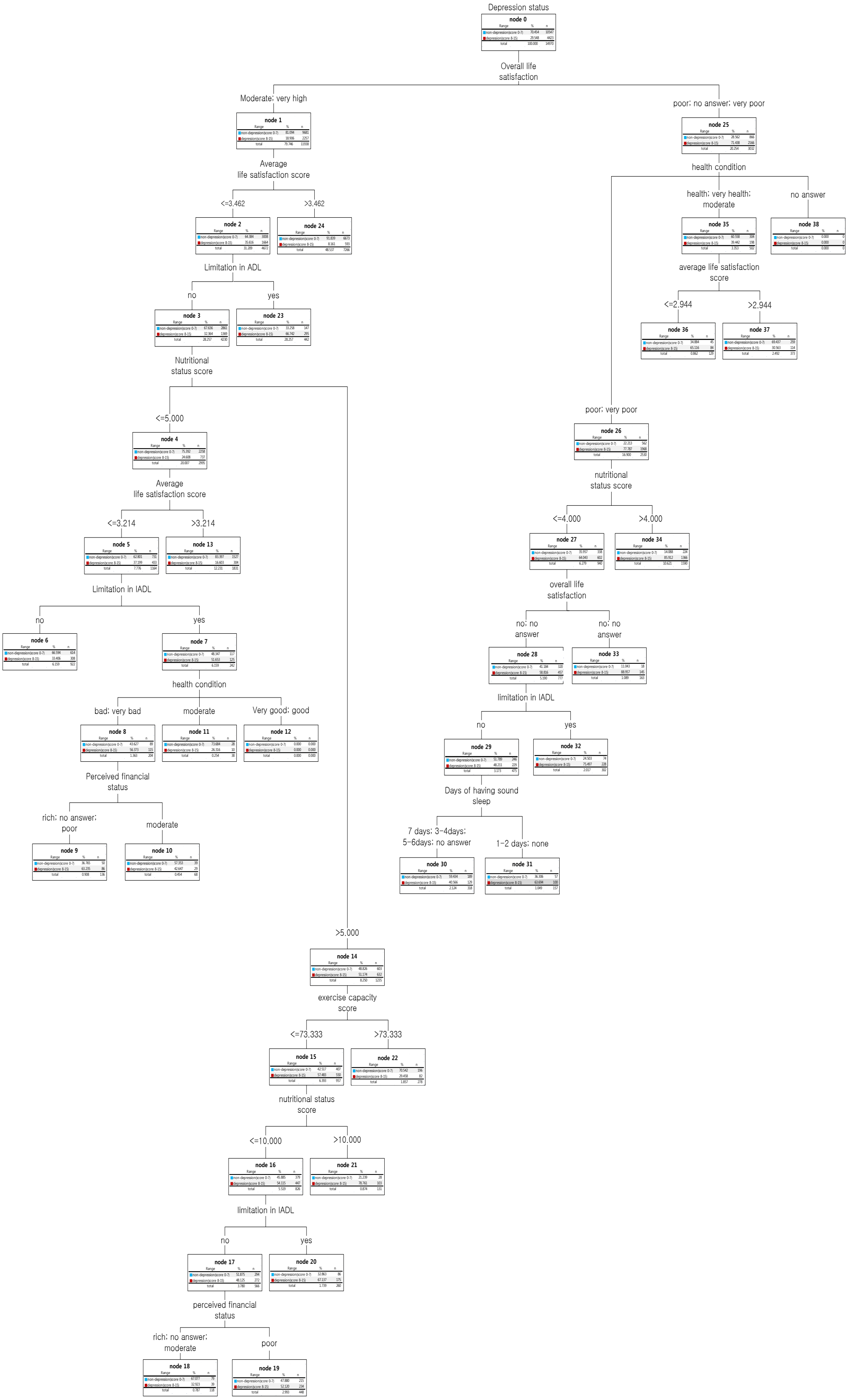

Figure 2. Decision Tree Model based on C5.0 Tree

We generated twenty rules from the full dataset and ten rules were associated with the depression. The ten decision rules were as follows: 1) moderate to high overall life satisfaction, average life satisfaction score was ≤3.462, having limitation in ADL; 2) moderate to high overall life satisfaction, average life satisfaction score was ≤3.462, no limitation in ADL, nutritional status score was between 5 and 10, exercise capacity score < 73.333; 3) moderate to high overall life satisfaction, average life satisfaction score was ≤ 3.462, no limitation in ADL, exercise capacity score <73.333, nutritional status score was > 10, no limitation in IADL; 4) moderate to high overall life satisfaction, average life satisfaction score was ≤ 3.462, no limitation in ADL, exercise capacity score <73.333, nutritional status score was > 10, having limitation in IADL, perceived financial status was poor; 5) moderate to high overall life satisfaction, average life satisfaction score was ≤3.462, no limitation in ADL, average life satisfaction score ≤3.214, having limitation in IADL, perceived health status was bad, perceived functional status was either rich or poor; 6) low overall life satisfaction, perceived health status was bad, nutritional status score was >4; 7) low overall life satisfaction, perceived health status was moderate to good, average life satisfaction score was ≤2.944; 8) low overall life satisfaction, perceived health status was bad, nutritional status score was ≤4, very low overall life satisfaction score; 9) low overall life satisfaction, perceived health status was bad, nutritional status score was ≤4, low overall life satisfaction score, having limitation in IADL; 10) low overall life satisfaction, perceived health status was bad, nutritional status score was ≤4, low overall life satisfaction score, no limitation in IADL, days of having sound sleep were 1-2 days per week or none.

4. Discussion

From a clinical point of view, one of the main issues in dealing with the elderly is to distinguish the elderly with depression. Using the C5.0 tree algorithm, 10 variables were selected including overall satisfaction of living, average life satisfaction score, limitation in ADL and IADL, nutritional status score, exercise capacity score, perceived health status and perceived financial status. The findings were consistent with previous studies that there were no single factor causing depression in elderly, in a range of socioeconomic, physical, psychological variables were closely associated with elderly depression [13].

Depression is a leading cause of low quality of life, therefore a method for timely detection should be ensured. The purpose of this study was to develop a decision support model based on statistical analysis and decision tree algorithm. Generally, the C5.0 Tree model showed good performance in predicting characteristics of depression in this study. Although, there are a variety of methods to analyze the clinical decisions, tree classification techniques have a few benefits over the alternative techniques [14]. The interpretation of results in a tree model can be useful not only for rapid classifying new clinical observations, but for explaining why observations are classified or predicted in a particular manner. In addition, performing data mining tasks in health care data, tree decision support models are particularly suitable as there is little priori knowledge that show which variables are related and how [14].

Ever increasing use of computer and health information system generates explosive amount of patient data. It is reported that those overwhelming data interfere with clinical judgment and decision making process in health care settings [15]. Data mining methods provides health professional with best solution to manage patient information and data. This study will provide researchers new insight into working with large healthcare datasets. We performed a 10-fold cross validation to estimate the accuracy of decision support models and 81.6% of accuracy was checked. As the results indicate, it is more effective to use larger induced decision model in distinguishing the characteristics of the elderly depression. The repeated testing and refinement of C5.0 tree modeling techniques with logistic regression in large healthcare databases, contributes to knowledge discovery and promote more sophisticated analyses.

5. Conclusion

This study developed a reliable decision support model to describe the characteristics of the elderly with depression performing statistical analyses and using a decision tree algorithm to provide high accuracy in decision making process. This model also facilitated discriminatory knowledge discovery in a large healthcare data set using the derived rules. The experimental results show that a decision tree model based on C 5.0 Tree with logistic regression provided excellent discrimination and we demonstrated its feasibility for classifying the elderly with depression. Therefore, the decision model developed in this study can be applied to identify the high risk group of depression and to increase vigilance when detecting depression in the elderly in the community.

Acknowledgements

This work was supported by Basic Science Research Program through National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0024922).

References

[1] S. C. Huang, E. C. Chang and H. H. Wu, “A case study of applying data mining techniques in an outfitter’s customer value analysis”, Expert Systems with Applications, vol. 36, no. 3, **(2009)**, pp. 5909-5915.

[2] L. C. Wu, J. X. Lee, H. D. Huang, B. J. Liu and J. T. Horng, “An expert system to predict protein thermostability using decision tree”, Expert Systems with Applications, vol. 36, no. 5, **(2009)**, pp. 9007-9014.

[3] S. W. Lin, Y. R. Shiue, S. C. Chen and H. M. Cheng, “Applying enhanced data mining approaches in predicting bank performance: A case of Taiwanese commercial banks”, Expert Systems with Applications, vol. 36, no. 9, **(2009)**, pp. 11543-11551.

[4] M. K. Obenshain, “Application of data mining techniques to healthcare data”, Infection Control and Hospital Epidemiology, vol. 25, no. 8, **(2004)**, pp. 690-695.

[5] J. Yu, J. Li, P. Cuijpers, S. Wu and Z. Wu, “Prevalence and correlate of depressive symptoms in Chinese older adults: A population-based study”, International Journal of Geriatric Psychiatry, vol. 27, **(2012)**, pp. 305-312.

[6] Y. H. Lee, M. H. Shin, S. S. Kweon, S. W. Choi, S. Y. Ryu, J. A. Rhee and J. S. Choi, “Prevalence and correlates of depression among the elderly in an urban community”, Journal of Agricultural Medicine & Community Health, vol. 33, **(2008)**, pp. 303-315.

[7] M. L. Steck, D. J. Vinkers, J. Gussekloo, R. C. Van Der Mast, A. T. Beekman and R. J. Westendorp, “Natural history of depression in the oldest old: Population-based prospective study”, British Journal of Psychiatry, vol. 188, **(2006)**, pp. 65- 69.

[8] M. Takeda and T. Tanaka, “Depression in elderly”, Geriatric & Gerontology International, vol. 10, **(2010)**, pp. 277-279.

[9] H. Svanstrom, C. Torbjorn and H. Anders, “Temporal data mining for adverse events following immunization in nationwide Danish healthcare databases”. Drug Safety, vol. 33, no. 11, **(2010)**, pp. 1015-1025.

[10] Ministry for Health and Welfare. 2008 Korean Elderly Survey. Seoul, **(2009)**.

[11] J. Huh, K. S. Jeong, S. H. Huh and H. K. Choi, Clementine 7 Manual, Data Solution, Seoul **(2003)**.

[12] J. R. Quinlan, “Induction of decision trees”, Machine Learning, vol. 1, **(1986)**, pp. 81-106.

[13] D. B. Kim and E. S. Sohn, “A meta-analysis of the variables related to depression in elderly”, Journal of the Korean Gerontological Society, vol. 25, no. 4, **(2005)**, pp. 167-187.

[14] L. Duan, W. N. Street and E. Xu, “Healthcare information systems: Data mining methods in the creation of a clinical recommender system”, Enterprise Information System, vol. 5, no, 2, **(2011)**, pp. 169-181.

[15] L. Goodwin, M. VanDyne, S. Lin and S. Talbert, "Data mining issues and opportunities for building nursing knowledge", Journal of Biomedical Informatics, vol. 36, no. 4, **(2003)**, pp. 379-388.

Authors

Myonghwa Park is an associate professor at college of nursing, Chungnam National University. Her research area includes data mining, health informatics, and evidence based nursing.

Chang Sik Son is a research scholar at Biomedical Informatics Technology Center, Keimyung University. His research area includes data mining, machine learning, and medical informatics.

Sun Kyung Kim is a doctoral student at college of nursing, Chungnam National University. Her research area includes nursing management and nursing informatics.