Fraudulent Bill-Claim Detection in Health Insurance

Junwoo Lee, Juhyeon Kim, Hyunjung Shin*

Abstract— **Fraudulent and abusive bill claims by medical care providers incur physical and fiscal costs to society. In order to identify them, a variety of indices have developed and evaluated diverse aspects of bill claim pattern. When taking all of indices into account, however, it becomes confusing to find out which index is of more importance than others, and even more difficult if the respective results are significantly discordant. To avoid the ambiguities, we propose a method that efficiently quantifies the degree of anomaly in the respective indices and then integrates them based on Genetic Algorithm. When tested on the Korean Health Insurance Review and Assessment data, the proposed method showed promising result of avg. 0.965 AUC, significantly outperforming the competing models including regression, neural network, and decision tree, etc.**

I. INTRODUCTION

Recently, the national medical expenses in South Korea is radically increasing. While the total expenses for medical treatment in 2002 had been 13,800 billion won, in 2007 it increased by 2.5 times and was 32,259 billion won, which shows a great upswing. The reasons for the increase include the development of medical facilities and the increase of elderly population, which are seen in natural aspects. On the other hand, they include negative aspects such as medical fraud. According to the report by NHCAA (National health Care Anti-fraud Association), it was assumed that 3~1-% of the total medical expenses in the United States (60 ~160 billion USD) were lost due to medical fraud. Judging from this result, we can expect that there would be a great scale of loss due to medical fraud in South Korea as well. Therefore, many studies on detection of medical fraud have been conducted to reduce the loss. In fact, however, there are many difficulties in terms of professional or technical aspects regarding a domain. First of all, the medical data requires professional understanding and its volume is tremendous. Because there is limited number of data processing professionals who have knowledge of a domain in South Korea, it is not easy to develop the appropriate system which deals with medical fraud. For the additional difficulty in terms of technical aspect, because the pattern of medical fraud is irregular, the fraud detection model generated from previous data is not suited for new data. For this reason, the previously developed detection system ends on the level of research and is rarely applied in practical setting. Therefore, the unjust or false claims have been detected manually by a few professionals. However, in reality we almost reach the limits in using the conventional method to detect the medical fraud which are getting developed over time while the data is greatly increasing.

The authors would like to gratefully acknowledge support from Post Brain Korea 21 and the research grant from National Research Foundation of Korean government (KRF-2009-0065043/ 2012-0000994).

* Corresponding author: Hyunjung Shin (shin@ajou.ac.kr) is a professor of the department of Industrial & Information Systems Engineering, Ajou University, Suwon, 443-749, Korea

Since the detection of medical fraud is a kind of detection of fraud, it is similar field to the detection of insurance fraud or credit card fraud. While the fraud detection system for insurance or credit card companies has been developed in response to the commercial requests of these companies, the fraud detection system in medical field has been developed based on academic research because medical field does not highlight commercial aspects compared to these companies.

According to the previous studies on medical fraud, the following methods were used: conventional statistical methods, data mining algorithms, and machine learning methods. The most mentioned methods in relevant studies are the Neural Network which exerts excellent effect on complicated data [4, 8] and the Decision Tree which is easily applied in practical setting because it makes interpretation of results easy [2]. The part of the studies conducted abroad was actually applied to the medical fraud system of each country and it showed high achievements compared to the manual labor done by a few professionals in the past. In Utah, the United States, they sort out the claim patterns which are suspected of unjust claims by analyzing data through data mining [9]. In Texas, the United States, they detected 1,400 fraud cases by using fraud detection system and collected 2.2 million USD [1]. The HIC of Australia separated meaningful rare data by applying genetic algorithm and k-NN algorithm [3,5]. It sorts out the claim patterns automatically, which was done manually by professionals [9]. The National Health Insurance (NHI) of Taiwan applies the clinical pathways to detect unknown unjust claims. The clinical pathway is a guideline for medical diagnosis and treatment that is defined by certain disease. Through this clinical pathway, they disclose the actions deviated from normal procedures for medical diagnosis and treatment [11].

The Health Insurance Review & Assessment Service (HIRA) of South Korea makes various claim indices and investigates the hospitals, clinics, dental clinics and oriental medicine clinics that are suspected of abuse or unjust use of medical expenses (hereafter, these institutions are called 'problematic institutions' in this paper). The current detection method has two problems by and large. The first problem is that the detection is not made based on the quantitative value, but made based on the order. If the judgment whether the institution is abnormal is made based on the order, it is impossible to show the difference between the problematic institutions and the rest of institutions by expressing numerically. It is difficult to show the level of severity. Therefore, the value of function which is based on the scores showing the level of abnormality, not the order, should be presented. The second problem is that the current detection method does not take all indices which are related to the indices of medical claim into consideration, but it puts weight only on single certain index. The single index cannot display the overall level of abnormality in each institution. In order to resolve these problems, the overall index was developed in the precedent study [6, 7] and it quantifies the level of abnormality of the medical claim pattern and unifies

individual index. It has been applied to the HIRA system since the second half of 2009. However, because the ideal scores suggested in the precedent study ranges quite widely, they have a tendency to expand the degree of abnormality. In addition, the calculation method for variance importance that was used in unifying individual index is to calculate a mean value of the weights obtained from several statistical analyses. In this study, therefore, we suggest the function made based on the precedent study but it can generate more sophisticated scores. Furthermore, in order to grant importance of variables, we suggest a methodology which uses genetic algorithm that is one of the meta-heuristic methods

II. Proposed Method

If the 'problematic institutions' is expressed in simple form in terms of medical claim, they means the institutions that show above average claim rates. Therefore, the important core concept in designing function is to focus the investigation on these institutions which show above average claims rates. In this study, we calculate the value of function in consideration of only the values which are higher than average by indices. By summing up these values according to the importance of the indexs, we divide the institutions into two groups: normal or problematic institutions.

A. Design of Scoring Function

Through the formula (1), the institutions which have above average value are given high value of function and the institutions which have below average value are given zero as the value of function.

$$P_i = \max\left(\frac{x_{ij} - \mu_j}{\sigma_j}, 0\right) \quad \textbf{(1)}$$

i means the record index and shows each institution of the data. j means the index of claim index. μ_j and σ_j means the mean and standard deviation of each claim index respectively. Therefore, zero is given as the value of function until the size of claim index reaches the mean. However, if it is larger than the mean, the higher value of function is given as the size of claim index becomes more distant from the mean, which is shown in Figure 1.

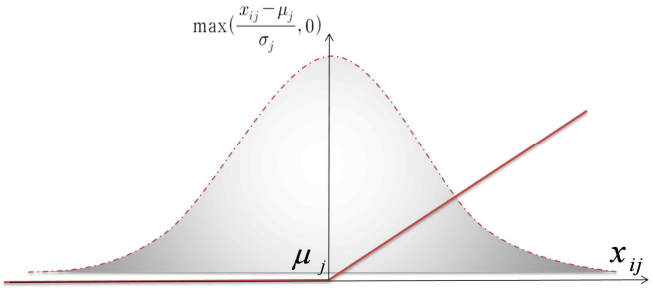


Fig.1. Function curve

Because the value of function for each claim index is obtained through the formula (1), it is necessary to sum up the values of function for all claim indices as shown in the formula (2) in order to reflect all claim indices.

$$S_i = \sum_{j=1}^{J} \alpha_j \max\left(\frac{x_{ij} - \mu_j}{\sigma_j}, 0\right) \quad \textbf{(2)}$$

The value of S_i is total value indices which presents the abnormality degree of institution and α_j is weight which presents the importance of each index (regarding the weight, we will address it in the following clause). With the value of S_i obtained through the formula (2), we design the score function as the formula (3) below in order to decide whether there is abnormality by using the critical value.

$$\hat{y}_i = \frac{2}{1+\exp\left[-\left(\alpha_0 + \sum_{j=1}^{J} \alpha_j \cdot \max\left(\frac{x_{ij} - \mu_j}{\sigma_j}, 0\right)\right)\right]} - 1 \quad \textbf{(3)}$$

The formula (3) is a sigmoid function and plays a role to make the value of $\hat{y}_i$ closer to binary variable by converting the value of S_i obtained through the formula (2) into the value which is located between -1 and 1 on the basis of the critical value (α_0). The obtained value of $\hat{y}_i$ is a discriminant score. If it is located close to -1, it means a normal institution. However, if it becomes closer to 1, it means an abnormal institution, a problematic institution. Figure 2 shows the division of abnormal institutions and normal institutions by the discriminant score on the basis of the critical value (α_0).

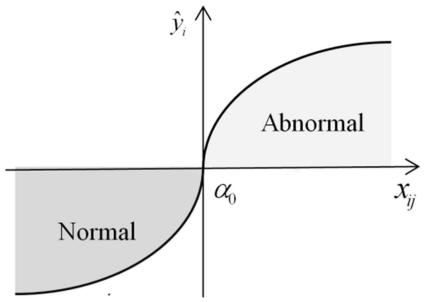


Fig. 2. Discriminant score for abnormal or normal institutions

B. Variable Weighting

In order to complete the formula (3) above, it is necessary to know the weight (α_j). At first sight, it seems that we might be able to obtain the weight by using the least-square method since the formula (3) is similar to the logistic regression analysis. However, from the formula (2), we can learn that this function cannot be differentiated. In other words, it is impossible to obtain the weight by using the least-square method. Therefore, we obtain the weight by using genetic algorithm (GA)[Davis, 1991; Holland, 1975; Goldberg, 1989]. GA performs the search process in four stages: initialization, selection, crossover, and mutation [Wong & Tan, 1994]. The initialization stage, a population of genetic structures, called chromosomes that are randomly distributed in the solution space, is selected as the starting point of the search. After the initialization stage, each chromosome is evaluated using a user-defined fitness function. The role of the fitness function is to numerically encode the performance of the chromosome. For real-world applications of optimization methods such as GAs, choosing the fitness function is the most critical step. In our study, the fitness function which is used in the space exploration process for resolution of genetic algorithm employs the Sum of Squared

Error (SSE) as shown in the formula (4). y_i is the actual value obtained from the data and $\hat{y}_i$ is the estimated value obtained from the formula (3). Under the fitness function of genetic algorithm, the chromosome which minimizes the formula (4) is chosen as the weight for variable.

$$\alpha = \arg\min \; f(\alpha) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \textbf{(4)}$$

The overall process of genetic algorithm is shown in Figure 3 and the algorithm is presented in Table 1.

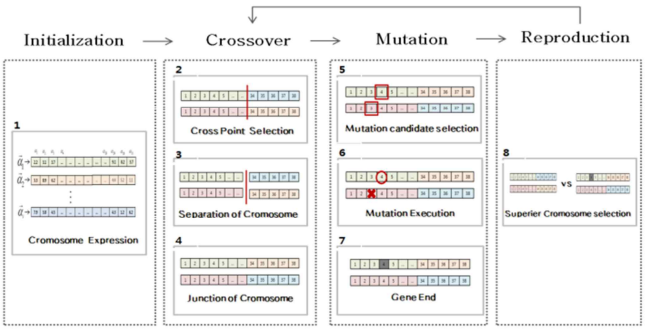


Fig.3. Genetic Algorithm application process for the exploration of variable weight

Table 1
Genetic Algorithm

Algorithm : Genetic Algorithm

begin initialize θ, P_{co}, P_{mu}, N_α, $\vec{\alpha}_i = [\alpha_1, \alpha_1, \cdots, \alpha_J]$, $\vec{\alpha}_i \in [0,10]$,

do determine fitness of each , $\vec{\alpha}_i$, f_i, $i = 1, \cdots, N_\alpha$

rank the $\vec{\alpha}_i$

do select two $\vec{\alpha}_i$ with highest score

if Rand[0,1] < P_{co} **then** crossover the pair at a randomly

else change each bit with probability P_{mu}

until N offspring have been created

until any $\vec{\alpha}_i$'s score f_i exceeds θ

return highest fitness $\vec{\alpha}_i$ (best α^*)

end

Each chromosome of genetic algorithm consists of the number of J genes which presents the importance of variable. N_α means the number of such gene. The genetic algorithm includes the process of reproduction, cross-breeding and mutation. Through these processes, the genes which have high fitness are chosen and the population is evolved. *θ, Pco*, and *Pm* means the critical acceptance value, crossover ratio, and mutation rate, respectively.

III. EXPERIMENT

The data used in the experiment was the claims data for medical expenses which was collected from the internal medicine clinics in Seoul area in the second half of 2007 and included the information of medical treatment institutions, details of claims, patients' information and information pertinent to claim settlement. The data consists of 600 which include 100 problematic institutions (y_i = 1) and 500 normal institutions (y_i =-1). The number of variable (claim index) is total 31. The value of the area under the ROC curve (AUC) was used to compare the predictive capability between methods [10]. The 5 fold Cross Validation (5-CV) was used for verification of model. With 5CV, the total data set is divided into five sets. Then four sets are used to make a model and the rest is used to verify the capability of the model. This procedure is repeated five times by applying the procedure to each set as much as possible. After repetition, the mean value is obtained as the final result value. In the following subsections, we first present the experimental results of the variable weighting with GA, and then the comparison results of the proposed method with the precedent method [6][7] and the representative data mining models.

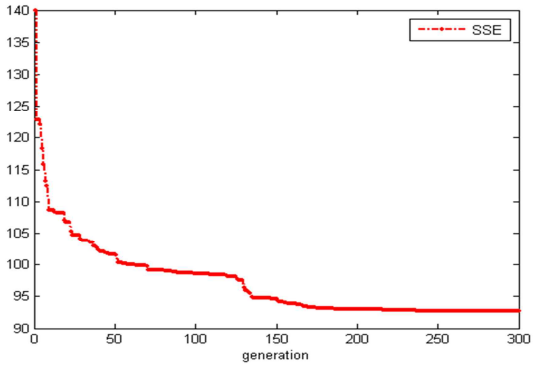
A. Variable Weighting with GA

The model parameter of genetic algorithm was set up as shown in the following Table 2.

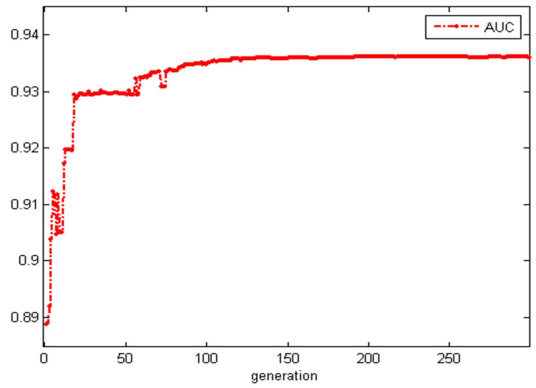
[Table 2]
Value of parameters of genetic algorithm

Parameter	
# of Pop	200
Prob. of crossover	0.80
Prob. of mutation	0.01
# of Generation	300

The genetic algorithm for established parameters explores the ideal weight by minimizing the fitness function which was designed in the previous clause. Because the weight which minimizes the fitness function maximizes the efficiency of a model, we can expect high accuracy. Figure 4 shows the convergent process of the value of fitness function (SSE) and the accuracy (AUC) as the generation proceeds. According to Figure 4(a), the SSE had continuously decreased as the generation proceeded. The SSE was above 140 at the first generation, but it decreased up to below 95 at the 300 generation. In addition, Figure 4(b) shows that the value of AUC continuously increased and that it increased from 0.89 to 0.94. This result shows that the accuracy improved as the generation proceeded.



(a) The value of fitness function (SSE) over the progress of generation

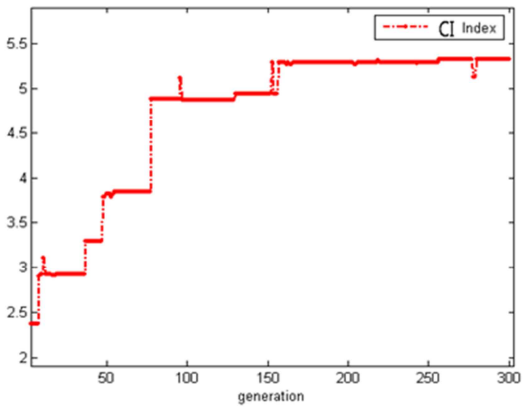


(b) AUC increase curve

Fig.4. Changes in the value of fitness function (SSE) and the accuracy (AUC) over the progress of generation

The importance of variables, in other words, the importance of assessment indices for claimed bills, is determined by the value of final chromosome gene which got through the evolution process of genetic algorithm. According to Figure 5, the weight of the index, 'CI[1)]' increased from 2.5 and then converged on near 5. The weight of the index, 'injection prescription rate,' started from 5.5. Then it continuously reduced and converged on near 2.5.

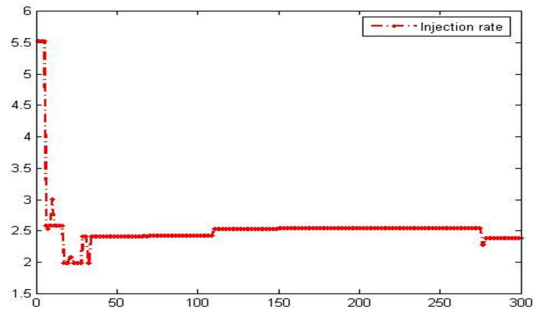
From Table 3, we can ascertain the relative value of weight of each of 31 indices which were obtained from the genetic algorithm. The size of the value means the importance of variable to explore the problematic institutions. Whereas the value of 'the visit day per case' was 1.09, it was 4.91 for 'Costliness index.' From this result, we can learn that 'Costliness index' has more important effect on exploring the problematic institutions by approximately 4~5 times than 'the visit day per case.



(a) Weight of CI

1) $$\text{Costliness Index} = \frac{\sum C_{hi} \cdot N_{hi}}{\sum C_i \cdot N_i}$$

(h: the institution, i: disease group)



(b)Weight of Injection prescription rate

Fig.5. Convergence graph of weights of CI and Injection prescription rate over the progress of generation

[Table 3]
Variable Weights
(*partial indices are not disclosed because of their confidentiality)

	Input Variables	MAD
1	Number of medicine	2.90
2	Costliness index	4.91
3	VI index	2.48
4	Medication expenses accrued outside the institution per treatment	3.04
5	Medication expenses accrued outside the institution per visit	2.29
6	CMI INDEX	1.55
7	Consultation fee CI	2.67
8	Oral administration CI	2.77
9	Psychological fees CI	2.96
10	Operation FEE CI	1.86
11	Diagnosis FEE CI	1.83
12	PET CI	1.64
13	Antibiotics prescription rate	2.01
14	Injection prescription rate	2.38
15	Medicine cost per	2.44
16	Rate of prescribed costly medicine	1.60
17	Number of medication per prescription	2.37
18	Rate of prescriptions more than 6 medicine	2.06
19	Digestives prescription ratio	1.48
20	Adrenalin Cortex-respiratory	1.30
21	Adrenalin Cortex-joint	0.76
22	Number of injury per detailed statement	2.50
23	Visit day per case	1.09
24	Administration day per case	2.55
25	Medication expenses accrued outside the institution per receiver	2.86
26	Total amount of treatment fees per receiver	2.43
27	top ranked CI	3.46
28	the 2nd ranked CI	3.10
29	the 3rd ranked CI	2.48
30	the 4th ranked CI	2.77
31	the 5th ranked CI	2.30

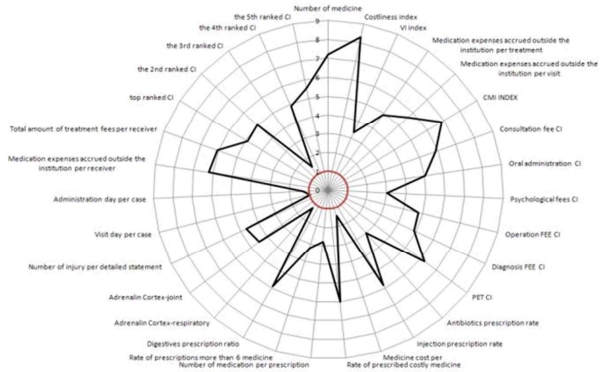


Fig.6. Diagram showing the relative importance of variable

B. *Comparison Results*

We compared the proposed method with the methods in the precedent study [6, 7] and several representative data mining algorithms. Hereafter, the method performing Medical Bill-Claim Abuser Detection is denoted as MAD for convenience. First, we compare the proposed method using the weight that was explored through the genetic algorithm (MAD_{GA}) and the method using equal weighting on the assumption that all α_j are the same (MAD_{EW}). The comparison experiment was conducted by carrying out 5-CV 15 times. The results were the mean AUC, which were shown in the second and the final rows of Table 4. From the comparison of the AUC which showed that the AUCs for MAD_{GA} and MAD_{EW} were 0.965 and 0.900, respectively, we learned that MAD_{GA} has much higher predictive accuracy. Moreover, from the comparison of the standard deviation which showed that the standard deviations for MAD_{GA} and MAD_{EW} were 0.017 and 0.008 respectively, we learned that MAD_{GA} is more stable. Since MAD_{GA} generates high accuracy and the changes in its results are stable, it is superior to MAD_{EW}.

[Table 4]
Performance Comparison with AUC Values
(MAD_{GA}, MAD_{EW}, and MAD_{PW} indicates different variable weighting methods, from GA (the proposed method in this study), from equal-weighting, and from the precedent study [6][7], respectively.)

15 5-CV	Reg	**NN**	**DT**	**MAD_{EW}**	**MAD_{PW}**	**MAD_{GA}**
1	0.923	0.907	0.669	0.912	0.933	0.968
2	0.920	0.918	0.630	0.893	0.925	0.956
3	0.885	0.746	0.759	0.904	0.922	0.968
4	0.916	0.919	0.825	0.860	0.912	0.955
5	0.888	0.906	0.646	0.924	0.903	0.982
6	0.924	0.913	0.795	0.905	0.901	0.968
7	0.900	0.904	0.749	0.896	0.936	0.970
8	0.919	0.900	0.646	0.886	0.916	0.951
9	0.862	0.908	0.669	0.918	0.922	0.968
10	0.913	0.930	0.763	0.894	0.917	0.961
11	0.882	0.901	0.693	0.899	0.880	0.972
12	0.900	0.920	0.773	0.917	0.937	0.966
13	0.871	0.881	0.749	0.898	0.922	0.975
14	0.901	0.879	0.798	0.918	0.938	0.964
15	0.886	0.897	0.749	0.879	0.929	0.958
μ	0.899	0.895	0.728	0.900	0.920	**0.965**
σ	0.020	0.044	0.063	0.017	0.016	**0.008**
p-value	**0.000** <0.05	**0.000** < 0.05	**0.000** < 0.05	**0.000** < 0.05	**0.000** < 0.05	

In the same way, we compared the weight obtained through the precedent study (MAD_{PW}) [6, 7] with the weight obtained through MAD_{GA}. The mean and standard deviation of AUC in MAD_{PW} were 0.920 and 0.016 respectively, which means that the efficiency of MAD_{PW} is short of the efficiency of MAD_{GA}. From the results comparison between 6th and 7th columns in Table 4, we can learn that the suggested genetic algorithm method is superior to the method obtained through the precedent study (MAD_{PW}).

The most prevalent methods for detection of medical fraud are Decision Tree (DT), Neural Network (NN), and regression analysis (Reg). The mean AUC for each method is shown in [Table 5]. While the means of AUC for regression analysis, neural network and decision tree were 0.899, 0.895 and 0.728 respectively, the AUC of MAD_{GA} was 0.965. Thus, the AUC of MAD_{GA} was highest among them. In addition, the standard deviation of AUC in MAD_{GA} was 0.008, which means that MAD_{GA} showed stable results compared to other algorithm. The very top of curve in Figure 7 was the ROC of MAD_{GA} and it always shows the highest value for all thresholding values. The regression analysis and neural network show the similar shape of the ROC curve. However, decision tree shows lower predictive value than other algorithms do as the thresholding value gets high while it shows high accuracy in low thresholding values.

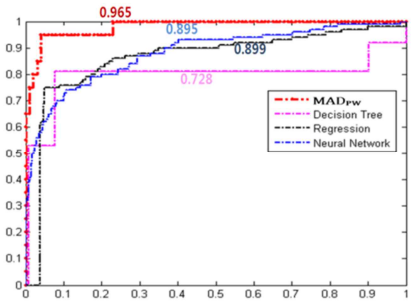
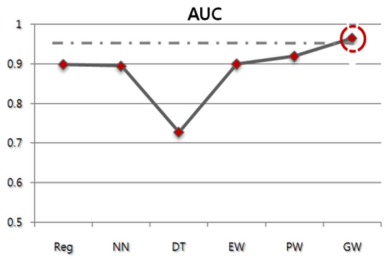
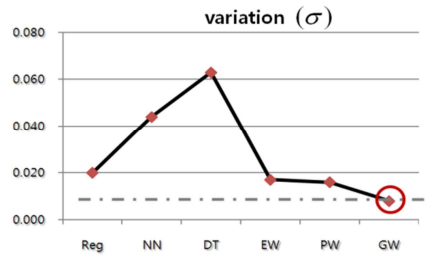


Fig.7. Comparison with ROC Curves

Figure 8 shows the mean and deviation of AUC values of different methods in Table 4. The two graphs, (a) and (b), present that the proposed method (MAD_{GA}) is the most accurate and stable than any other methods in comparison



(a) Comparison of AUC



(b) Comparison of variation

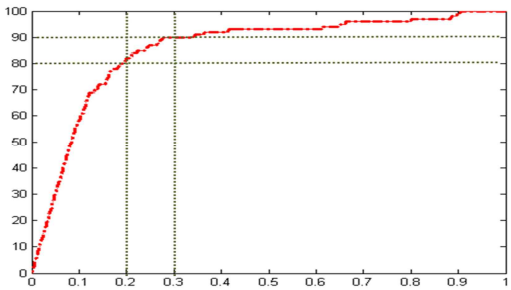
Fig.8. Comparison of accuracy and stability.

The last two rows of Table 6 present the result of t-test which tested whether there was a significant difference between MAD_{GA} and other methods. Through t-test, we can test whether the difference between AUC of other methods and that of MAD_{GA} is caused by accident or by the actual difference in efficiency. Because the p-values were near 0 which was lower than 0.05, a significance level, we can learn that the excellent efficiency of MAD_{GA} is statistically significant.

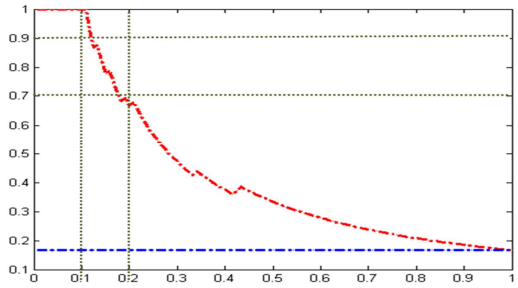
C. *Practical Implication*

Figure 9 shows Lift and Response chart which were generated through MAD_{GA}. The Y-axis in Lift chart means the number of problematic institution. In Response chart, it means the hit ratio of problematic institution. The x-axes of two charts mean the institutions which were arrayed based on the predictive values obtained from the formula. According to Lift chart in Figure 9(a), if we select the top 20% problematic institutions, we can explore 80% of the total problematic institutions. If the top 30% are selected, 90% can be explored. From Response chart in Figure 9(b), if the top 10% problematic institutions are selected, all of them are problematic institutions. If the top 20% problematic institutions are selected, 70% of the selected institutions are actual problematic institutions. After reviewing these two charts, we can interpret that if we selected 20% of institutions based on the predictive value of the suggested model, we are able to find 80% of the actual problematic institutions with 70% accuracy.

In practical setting, time and manpower generate costs. Therefore, if we use the suggested method which has high accuracy, we can expect the effect of cost reduction and improvement of efficiency because we can explore many unjust institutions with a little data.



(a) Lift chart



(b) Response Chart

Fig.9. Lift and Response charts

IV. Conclusion

In this study, we developed the medical fraud detection model by using the data in relation to medical claim for the purpose of establishing efficient national just medical expenses review system. Through the score function which is one of the suggested methods and measures the degree of abuse and unjust use of medical expenses made by health institutions focuses only on a few institutions which have a value above the mean of medical claim expenses made by all health institutions, excluding the institutions which have a value below the mean. By doing this, the score function was designed to effectively work in terms of calculation and actual applicability. Furthermore, for the determination of weight which decides the importance of the claim indices (variables) that are in relation to medical expenses overuse or abuse, we introduced the genetic algorithm. Although this method is simplified compared to the methods obtained through the precedent study [6, 7], it provide much more accurate methodology. The suggested method was compared with the precedent study, decision tree, neural network and regression analysis. The results show that the suggested model is a stable model with high predictability.

In the actual process of medical expenses review, after the problematic institutions which claims unjust or false medical expenses are selected, the affected institutions always raise issues or resist during the course of punishment procedure. Therefore, the model for selection must be accurate and the reason for selection should be clearly presented. The model suggested in this study has high predictability. Besides, the variable weight method using the genetic algorithm makes it clear which variable is important in exploring the problematic institutions and how much important the variable is. Thus, this method provides efficiency by raising the transparency in relation to the results of medical expenses review.

The establishment of efficient and effective medical fraud detection system by HIRA not only saves the medical expenses but also uses medical insurances correctly which all people have to shoulder in terms of social expenses. Therefore, we will develop much more accurate and efficient medical detection system by introducing the latest data mining method and machine learning method in the further study.

Acknowledgment

H.Shin would like to gratefully acknowledge support from Post Brain Korea 21 and the research grant from National Research

Foundation of the Korean Government (KRF-2009-0065043 /2012-0000994)

REFERENCES

[1] D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*. MA: Addison-Wesley, 1989.

[2] H. J.H., *Adaptation in Natural and Artificial System: An Introduction with Application to Biology*. Ann Arbor, MI: University of Michigan, 1975.

[3] D. L., *Handbook of Genetic Algorithms*. New York, NY: Van Nostrand Reinhold, 1991.

[4] P. L. Bartlett*, et al.*, "Learning Changing Concepts by Exploiting the Structure of Change," *Machine Learning,* vol. 41, 2000.

[5] F. Bonchi*, et al.*, "A classification-based methodology for planning audit strategies in fraud detection," in *ACM SIGKDD*, New York, NY, USA, 1999.

[6] H. He*, et al.*, "Application of genetic algorithms and k-nearst neighbour method in medical fraud detection," in *SEAL*, Springer-Verlag London, UK, 1999.

[7] H. He*, et al.*, "Application of neural networks to detection of medical fraud," *Lecture Notes in Computer Science,* vol. 1585, pp. 74-81, 1999.

[8] H. Shin and J. Lee, "How to Integrate the diverse measures for hospital fraud detection," in *INFORMS*, San diego, CA, USA, 2009.

[9] H. Shin*, et al.*, "A scoring model to detect abusive billing patterns in health insurance claims," *Expert Systems with Applications,* vol. 39, pp. 7441-7450, 2012.

[10] J. A. Major and D. R. Riedinger, "EFD: A hybrid knowledge/statistical-based system for the detection of fraud," *Risk and Insurance,* vol. 69, pp. 309-324, 2002.

[11] G. J. Williams and Z. Huang, "Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases," *Lecture Notes in Computer Science,* 1997.

[12] V. M. and C. T., "ROC curve, lift chart and calibration plot," *AMS,* vol. 3, pp. 89-108, 2006.

[13] W.-S. Yang and S.-Y. Hwang, "A process-mining framework for the detection of healthcare fraud and abuse," *Expert Systems with Applications,* vol. 31, pp. 56-68, 2006.