Original Article

Healthc Inform Res. 2010 June;16(2):67-76.
doi: 10.4258/hir.2010.16.2.67
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Classification and Sequential Pattern Analysis for Improving Managerial Efficiency and Providing Better Medical Service in Public Healthcare Centers

Keunho Choi, PhD[1], Sukhoon Chung, PhD[2], Hyunsill Rhee, PhD[3], Yongmoo Suh, PhD[1]

[1]Business School, Korea University; [2]Korean Institute of Hospital Management; [3]Department of Healthcare Management, College of Health Science, Korea University, Seoul, Korea

Objectives: This study sought to find answers to the following questions: 1) Can we predict whether a patient will revisit a healthcare center? 2) Can we anticipate diseases of patients who revisit the center? **Methods:** For the first question, we applied 5 classification algorithms (decision tree, artificial neural network, logistic regression, Bayesian networks, and Naïve Bayes) and the stacking-bagging method for building classification models. To solve the second question, we performed sequential pattern analysis. **Results:** We determined: 1) In general, the most influential variables which impact whether a patient of a public healthcare center will revisit it or not are *personal burden, insurance bill, period of prescription, age, systolic pressure, name of disease, and postal code*. 2) The best plain classification model is dependent on the dataset. 3) Based on average of classification accuracy, the proposed stacking-bagging method outperformed all traditional classification models and our sequential pattern analysis revealed 16 sequential patterns. **Conclusions:** Classification models and sequential patterns can help public healthcare centers plan and implement healthcare service programs and businesses that are more appropriate to local residents, encouraging them to revisit public health centers.

Keywords: Public Healthcare Center, Data Mining, Classification Analysis, Sequential Pattern Analysis, Ensemble Method

Received for review: February 16, 2010
Accepted for publication: May 12, 2010

Corresponding Author
Keunho Choi, PhD
Business School, Korea University, Anam-dong 5-ga, Seongbuk-gu, Seoul 136-701, Korea. Tel: +82-2-3290-1945, Fax: +82-2-922-7220, E-mail: keunho@korea.ac.kr

I. Introduction

Promoting and maintaining a good public health is a growing concern both of national and of regional governments of Korea as in other countries, and it contains activities which contribute to developing public health policy and delivering healthcare services [1]. Public health sector should take over and manage the healthcare services which have been ignored by private health sector, and this viewpoint should be reflected in public health policy [2].

In 1995, Korean government enacted a law on public health promotion, where it changed its view that public health management of local residents should be carried out not by central public healthcare centers but by local ones. Korea public healthcare centers offer various programs and ser-

vices to local resident, such as those about quitting smoking, moderation in drink, nutrition, hypertension, diabetes, arthritis, and cancer etc. However, operations of most public healthcare centers do not consort with local residents' demand [3] and therefore, their managerial efficiency is known to be very low [4].

Moreover, in spite of such political movement by government, health and welfare services are under the leadership of private health sector, and public health sector still remains suffering from identity crisis [2]. According to Nam's study [5], 80% of local residents recognized the necessity of public healthcare centers; 39.2% of them thought that the most important function of public healthcare center was medical treatment; 40.2% of them hesitated to visit public healthcare centers because they worry about the lack of public health center staffs' expertise; 38.4% of them were dissatisfied most with the lack of promotional activities.

Public healthcare centers need to make an effort to resolve these problems. At the macro level, for improvement in managerial efficiency, it is necessary for public healthcare centers to establish and execute flexible policy that reflects health demand by local residents and allocates budget and manpower according to the health demand [3]. At the micro level, the local governments have to do their best for health improvement of the local community by planning and implementing healthcare service programs and businesses, encouraging local residents to visit public health centers and guaranteeing sufficient healthcare resources and so on [5]. Since the local healthcare businesses are conducted by public healthcare centers, it is important to have public healthcare centers play their roles and perform their functions well [5].

Public healthcare centers aim to offer substantial healthcare services to local residents by establishing healthcare policy suitable for local conditions. Thus, plans for this healthcare policy (*e.g.*, organization, manpower, facilities, equipments, budget, and promotion plans) can be made through the estimation of health demand, which can be estimated by predicting patient's revisit as a starting point. Therefore, preferentially, knowing about the possibility of patient' revisit is essential to establishing suitable healthcare policy, appropriate to local conditions.

In addition, public healthcare centers make an effort to provide information about methods for disease prevention and treatment with local residents. According to the report of National Health Insurance Corporation, diseases such as essential hypertension, acute upper respiratory infection, diabetes, tooth and supporting structure trouble, soft tissue trouble, rheumatoid arthritis and arthropathy, gastritis and duodenitis, dental caries, endocrine and metabolic diseases, and skin and tissue trouble are the most frequent diseases of patients who visited public healthcare centers from 2003 to 2005. To provide better medical services in terms of prevention and treatment of most frequent diseases, it is required to inform patients of precautionary measures according to their foreseeable diseases.

Having discussed with the staffs of a public healthcare center located in the north of Seoul, Korea, we decided to analyze the center's patients data accumulated from January 1, 2007 till June 24, 2008, with a purpose to find answers to the following research questions: 1) Can we predict whether a patient will revisit the public healthcare center?; 2) Can we suggest foreseeable diseases to patients who revisit the center? Answers to these questions will be helpful in improving managerial efficiency of the center and providing a better medical service to the patients by suggesting precautionary measure to them.

As a means to provide an answer to the first question, we first applied five classification algorithms (*i.e.*, decision tree [DT], artificial neural network [ANN], logistic regression [LR], Bayesian networks [BN], and Naïve Bayes [NB]) and stacking-bagging (SB) method proposed in this study to building classification models. To solve the second research question, we performed sequential pattern analysis with a purpose to identify foreseeable diseases of revisiting patients. All the details about these experiments are given later in section 3.

The rest of this paper is organized as follows. In section 2, we reviewed previous researches which make use of classification or sequential pattern analysis in medical or healthcare domain. Section 3 describes our research method, including description of data, overall structure of our experiments, pre-processing and variable selection, and the experiments we carried out. Section 4 explains the results of our experiments and compares them. In section 5, we conclude the paper with a summary, implication of the research results and limitations.

1. Literature Review

Increasing use of data mining techniques can be found in a wide variety of areas such as finance, retail markets, telecommunication, medical area, and so on [6-10]. Previous research has shown that data mining techniques can be used to elicit untapped useful knowledge from large medical datasets [11,12]. This section reviews previous research which utilizes *classification or sequential pattern analysis* for various tasks in medical or healthcare domain.

Classification tasks have been carried out for various purposes in medical or healthcare domain. Choi et al. [13] pro-

 doi: 10.4258/hir.2010.16.2.67

posed a hybrid model by combining the artificial neural network and Bayesian network to predict 5-year survival rates for breast cancer. Diri and Albayrak [14] adopted Bayesian network, k-nearest neighbor, k-means, and self-organizing map to classify the thyroid gland patients into 3 classes (hyperthyroidism, hypothyroidism, and euthyroidism class). Phillips-Wren et al. [15] used 3 data mining techniques – decision tree, logistic regression, and artificial neural network – to predict whether a lung cancer patient will visit the medical oncologist or not. Chang and Chen [16] used decision tree and artificial neural network to predict 6 types of skin diseases in dermatology. Lee and Shih [17] investigated the potential of artificial neural network in recognizing profitable customers for the operation of dental clinics, and compared accuracy of artificial neural network with that of discriminant analysis. Polat et al. [18] used decision tree in order to classify healthy and macular diseased subjects. Ture et al. [19] used 6 decision tree algorithms – classification and regression tree, chi-squared automatic interaction detector, quick, unbiased and efficient statistical tree, iterative dichotomiser 3, commercial 4.5 (C4.5), commercial 5.0 and cox regression to predict the disease-free survival in breast cancer patients. Wu et al. [9] in their study adopted Naïve Bayes, decision tree, and artificial neural network to develop a predictive model for protein thermostability based on sequence and structural features. Chang et al. [20] implemented a support vector machine based system to automatically identify the health related information on the webs. Kang et al. [21] developed 2 artificial neural network models and 2 classification and regression tree to predict both the total amount of hospital charges and the amount of expenses paid by the insurance of cancer patients.

Another data mining technique that is useful for the analysis of medical or healthcare data is sequential pattern analysis, for one disease may be progressed into another in many cases. Exarchos et al. [22] analyzed protein sequence and classify proteins into the folds. That is, they extracted sequential patterns of proteins, which were then used to classify the unknown proteins. Chiang et al. [23] extracted interaction patterns between genes obtained from biomedical documents. To be specific, this study developed a new sequential pattern mining method to mine meaningful rules that describe the kinds of morphological features that can appear before and after the name of gene in documents. Ryan [24] examined sequences of health-related behaviors from a small village in Cameroon. One of the findings from their study is that residents' first use delay of treatments as a strategy in the decision making process, then rely on home-based treatments and then seek treatment from outside the compound. Lasker [25] identified patients' disease on the basis of their sequential symptoms. When patients have one of the foreseeable diseases, each of these diseases may be confirmed by specific sequential symptoms. Lin et al. [26] developed a sequential data mining technique which is helpful to organize patient care activities, to diminish practice variations, and to minimize delays in treatments for the purpose of facilitating the continuous improvement of assigning more suitable clinical paths to brain stroke patients. Concaro et al. [27] exploited sequential pattern analysis to discover frequent sequential and association patterns of diagnoses shared by United States hospitals. In addition, sequential patterns identified can provide a descriptive scenario of the temporal advance of the most frequent healthcare episodes during the year. On the other hand, association patterns not only describe sets of synchronized event, but also suggest potential associations between involved diseases.

From the literature review, it can be seen that more and more analysis of medical or healthcare data are analyzed using data mining techniques for classification or prediction tasks to derive knowledge that can be used for decision making in medical or healthcare domain. In this study, we also applied classification algorithms and ensemble techniques to building a model to classify patients of a public healthcare center into re-visitor or into one-time visitor, and applied sequential pattern analysis technique to identifying foreseeable diseases of revisiting patients.

II. Methods

1. Data

The data used in this study were provided by a public healthcare center of Korea, after removing confidential information. The original database of public healthcare center contains 20 relations which include such tables as resident master, application, receipt, prescription, judgment, blood pressure, vaccination, pregnant, child, and so on. The entire data covers the period from January 1, 2007 till June 24, 2008 (18 months) and includes 39,388 instances. We eliminated duplicate samples or those with many missing values, and finally we obtained remaining 7,057 samples.

2. Research Architecture

Our research architecture shows 2 streams of research activities. One is to build classification models of revisiting patients and the other is to find sequential patterns of diseases. Prior to the classification task, we preprocessed the data, and selected variables to be used in our study. Then, in the first stage of classification task, we used 5 classification

techniques such as decision tree, artificial neural networks, logistic regression, Bayesian networks, and Naïve Bayes to build plain classification models and compared them based on the results obtained from cross-validation. In the second stage of classification task, we applied both stacking and bagging techniques to the base classification models to get more reliable results. And then, we compared the classification accuracy of the plain models obtained in the first stage with that of the ensemble model obtained in the second stage, to find a classification technique most suitable for predicting patient' revisit. On the other hand, for sequential pattern analysis, we preprocessed the data, and utilized sequential pattern mining technique to find sequential patterns among the diseases of revisiting patients. After verifying the sequential patterns, we obtained meaningful ones that can be used to predict foreseeable diseases of revisiting patients and then to provide them with adequate precautionary measures.

3. Experiments for Classification Task

1) Preprocessing

We examined whether a patient revisited the public healthcare center in 3, 6 or 12 months after his or her first visit to create a target field, revisit, which is a Boolean variable. Therefore, 3 datasets were prepared from the original dataset to build 3 classification models, one to tell whether a patient will revisit in 3 months (called 3M dataset from now on), another in 6 months (6M dataset), and the third in 12 months (12M dataset), respectively. 3M dataset includes 1,464, 6M dataset 1,289, and 12M dataset 1,001 instances, with duplicate revisiting patient data being deleted. The portion of revisiting patients in 3M, 6M, and 12M datasets accounted for 41.12%, 50.04%, and 67.03%, respectively. Finally, we adjusted the 3M (1,204 instances) datasets by under-sampling and 12M (1,342 instances) datasets by over-sampling so that the portion of revisiting patients becomes almost 50%, as in 6M dataset.

2) Variable selection

Many variables have been used in the previous researches to predict whether patients will revisit a public healthcare center or not. Phillips-Wren et al. [15] used both socio-demographic and clinical characteristic data in their study to predict whether a lung cancer patient will visit the medical oncologist or not. These variables indicate patient conditions, demographics, and treatments. They have been validated in various healthcare-related studies [15,28,29]. *Distance* and *treatment cost* also are reported that they also affect the possibility of visit [30-32].

For our classification task, we have discussed with the staffs of a public healthcare center and reviewed previous healthcare-related studies [15,30-34]. Through the above discussion and literature review, we chose 10 input variables (*i.e.*, gender, age, zip code, insurance bill, personal burden, insurance type, period of prescription, name of disease, systolic blood pressure, and diastolic blood pressure) from 5 tables of our database, and added a new variable, distance, derived from patient' address. As a target variable, we used patient's revisit. Table 1 shows description of variables for classification analysis. We took the wrapper approach to decide a final set of variables with stepwise backward elimination, while each plain classification algorithm was used to evaluate each set of variables, which is explained further in the next section.

3) Plain classification models

To conduct our classification task, we used Weka ver. 3.6 (open source software) as a data mining tool, which is widely used for various data analysis. We evaluated 11 input variables using Gain Ratio attribute evaluator based on ranker search method, to select more influential variables when predicting the target variable. Table 2 shows the importance ranking of input variables in each dataset. Personal burden,

Table 1. Variables used for classification

Variable	Description
Gender	Male, female
Age	Patient's age in number
(Derived) Distance	Distance between public healthcare center and patient' address
Zip code	Postal code
Insurance bill	Medical expense covered by insurance (Korean won)
Personal burden	Patients' share in medical expense (Korean won)
Insurance type	Type of insurance
Period of prescription	Days for which doctor's prescription is valid
Name of disease	Code (indicating the name of disease)
Systolic pressure	The highest arterial pressure during heart beat (mmHg)
Diastolic pressure	The lowest arterial pressure during heart beat (mmHg)
Revisit	Whether a patient revisits, or not

doi: 10.4258/hir.2010.16.2.67

Table 2. Ranking of attributes

Attributes	Ranking			Attributes	Ranking		
	3M	6M	12M		3M	6M	12M
Personal burden	1	1	1	Diastolic pressure	7	8	9
Insurance bill	2	2	2	Zip code	8	7	7
Period of prescription	3	3	3	Distance	9	10	8
Age	4	5	5	Insurance type	10	9	10
Systolic pressure	5	4	6	Gender	11	11	11
Name of disease	6	6	4				

3M: 3 months dataset, 6M: 6 months dataset, 12M: 12 months dataset.

insurance bill and days of prescription are top 3 influential input variables, age, name of disease and systolic are the next top 3 influential input variables, while gender is the least influential input variable to predict patient's revisit in all 3 models.

With the ordered list of input variables, we adopted wrapper approach, in which stepwise backward elimination method is used to select a proper subset of attributes for each of 5 different classification algorithms such as decision tree, artificial neural network, logistic regression, Bayesian network, and Naïve Bayes. To build a decision tree model, we used C4.5 algorithm which has showed good performance in previous researches. The parameters of artificial neural network such as learning rate, momentum, epoch, and the number of hidden-layer were set to 0.3, 0.2, 50, and 1, respectively. The parameters of the other classification algorithms were set to the default values in Weka. Having built classification models using the 5 classification techniques, we evaluated and compared their classification results obtained from 5-fold cross-validation. Experimental results are described in Section 4.

4) Ensemble classification model

With the 5, 6, and 7 variables selected as a result of building plain models in 3M, 6M, and 12M dataset, respectively, we then applied both stacking and bagging techniques in a row to the best 4 base classifiers after removing the worst base classifier in each dataset to get more reliable results. Both stacking and bagging are ensemble approaches which combine the results of multiple classifiers. In general, ensemble classifiers have been reported to result in better classification than a plain classifier [35-37]. The rationale of ensemble classifier is that making a decision after combining the results of several classifiers would be better than making a decision solely based on a single classifier, as we ask for the opinions of several doctors before undergoing a serious surgical operation. Stacking and bagging are different in some aspects. Their first difference is that bagging uses multiple classifiers of same type, while stacking uses multiple classifiers of different type. Another difference between them is that bagging combines models built from multiple training datasets each of which is obtained by sampling with replacement from a single dataset, while stacking combines models built from solely one training dataset. With the expectation that we can build a more reliable classification model, we proposed stacking-bagging method which is an ensemble of ensembles. In order to combine results from the plain classifiers, we used the majority voting which has been adopted in ensemble models generally.

The dataset representing one of 3M, 6M, and 12M is used for 5-fold cross validation, as we did to build plain models. That is, dataset is partitioned into 5 sub-datasets, and one is reserved to be used as a test dataset, while the rest is used to build an ensemble model. Another sub-dataset is then reserved as a test dataset and the rest is used to build another ensemble model. This repeats 5 times. Experimental results are given in Section 4.

4. Experiments for Sequential Pattern Analysis

1) Preprocessing

Since we want to find sequential patterns that occur during the whole period (If we have a huge amount of patient data, it would be better to change the size of the time window for sequential patterns to X month or to Y year, dynamically) of dataset (*i.e.*, 18 months, from January 1, 2007 till June 24, 2008), we arranged duplicated patient-IDs in ascending order of their date of visit. Each record contains the name of disease which patients have on the date of visit. Since our dataset contains many patients who have only one disease, the

support value of meaningful sequential patterns with two or more length to be found may go down below the minimum support. To find more latent sequential patterns which may not be found when considering all patients including those who have only one disease, we made 3 sequence datasets including patients who have more than 2 (named sequence dataset 1), 3 (named sequence dataset 2), and 4 (named sequence dataset 3) individual diseases. Sequence datasets 1, 2, and 3 contain 326 (817 transactions), 114 (393 transactions), and 32 (147 transactions) instances, respectively.

2) Parameters for sequential pattern analysis

To conduct our sequential pattern analysis, we used SAS ver. 9.1 Enterprise Miner (SAS Institute Inc., Cary, NC, USA). As mentioned above, we used patient-ID, date of visit, name of disease as ID, sequence, and target variables, respectively. Because of the small number of transaction for each patient-ID, we set time window to be unlimited in order to find sequential patterns that may span the whole period of the dataset. In sequence dataset 1, 2, and 3, we set the minimum support to be 1%, 3%, and 6%, and the minimum confidence to be 10%, 15%, 35%, respectively.

III. Results

1. Results from Plain Classification Models

Eleven experiments for each of the 5 data mining techniques, or a total 55 experiments in each dataset were conducted for classification analysis. Figure 1 depicts the average of classification accuracy obtained from each data mining technique with stepwise backward feature elimination in 3M, 6M, and 12M datasets. From this figure showing the fluctuation of classification accuracy as the number of input variables decrease, we can see that the highest classification accuracy was acquired when we used the most influential 5 to 7 variables in most cases for our data sets.

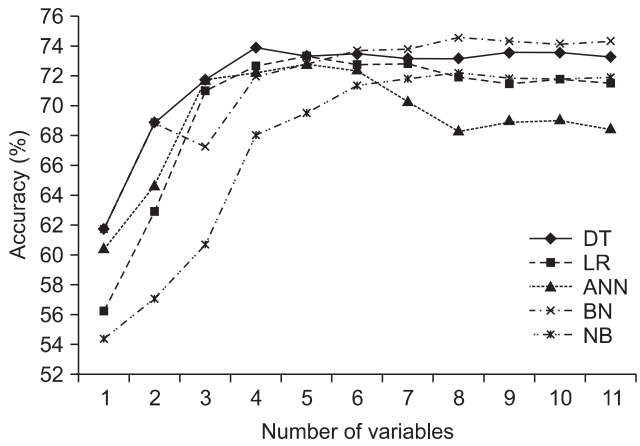

Figure 1. Average of classification accuracy of each data mining techniques. DT: decision tree, LR: logistic regression, ANN: artificial neural network, BN: Bayesian networks, NB: Naïve Bayes.

For the experiments with 3M dataset, the best classification accuracy (72.84%) was achieved by artificial neural network with first 5 input variables. Decision tree, logistic regression, Bayesian network, and Naïve Bayes show their highest classification accuracy (71.84%, 71.84%, 72.51%, and 70.85%, respectively) with the first 4 (or 6), 5, 8 (or 9), and 8 variables, respectively.

For the experiments with 6M dataset, the best classification accuracy (73.03%) was achieved by decision tree with the first 9 or 10 input variables. Logistic regression, artificial neural network, Bayesian network, and Naïve Bayes show their highest classification accuracy (72.69%, 72.92%, 72.85%, and 71.22%, respectively) with the first 5, 3, 8 and 9 variables, respectively.

For the experiments with 12M dataset, the best classification accuracy (78.69%) was achieved by logistic regression with the first 7 input variables. Decision tree, artificial neural network, bayesian network, and Naïve Bayes show their highest classification accuracy (77.42%, 76.38%, 78.32%, and 75.19%, respectively) with the first 5, 6, 8, and 8 variables, respectively.

From the results of plain classification models, we can see that generally most data mining techniques achieve their best performance with first 5, 6, and 7 variables in 3M, 6M, and 12M dataset, respectively. Therefore, we used these variables in each dataset to conduct further experiments. In all datasets, classification models maintain their classification accuracy to some extent as the number of input variables increases except logistic regression and artificial neural network. To build the stacking-bagging method, the best 4 plain classifiers were used after removing the worst plain classifier - Naïve Bayes in 3M dataset and artificial neural network in both 6M and 12M datasets - for better performance.

2. Results from Ensemble Classification Model

As shown in Table 3, stacking-bagging method proposed in this study outperformed all the best plain techniques - artificial neural network in 3M dataset, decision tree in 6M dataset, and logistic regression in 12M dataset. In addition, stacking-bagging method also outperformed the bagging of each best plain technique, and stacking in all three datasets except only stacking in 12M dataset. Although stacking method outperformed stacking-bagging method a little bit only in 12M dataset, we can see that stacking-bagging meth-

 doi: 10.4258/hir.2010.16.2.67

Table 3. Classification accuracy of plain best, bagging of each best plain technique, stacking, and stacking-bagging method in each dataset

Dataset	Plain best	Bagging					Stacking	Stacking-bagging
		DT	LR	ANN	BN	NB		
3M	72.84	70.60	72.01	72.26	72.18	70.60	72.59	72.92
6M	73.03	71.76	72.85	72.46	72.92	69.67	71.68	74.17
12M	78.69	77.20	78.39	77.42	77.79	75.78	78.91	78.84
Average	74.87	73.19	74.42	74.05	74.30	72.02	74.39	75.31

Values are presented as percent.

DT: decision tree, LR: logistic regression, ANN: artificial neural network, BN: Bayesian networks, NB: Naïve Bayes, 3M: 3 months dataset, 6M: 6 months dataset, 12M: 12 months dataset.

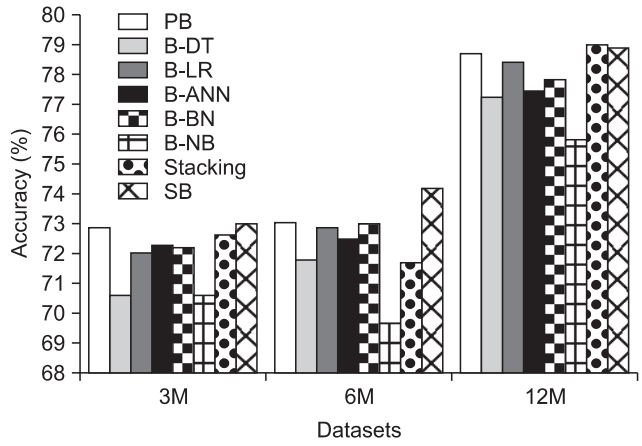

Figure 2. Classification accuracy of plain best, bagging of each best plain technique, stacking, and stacking-bagging method in each dataset. PB: plain best, B-DT: bagging of decision tree, B-LR: bagging of logistic regression, B-ANN: bagging of artificial neural network, B-BN: bagging of Bayesian network, B-NB: bagging of Naïve Bayes, SB: stacking-bagging. 3M: 3 months dataset, 6M: 6 months dataset, 12M: 12 months dataset.

od can give better performance (75.31%) than the best plain technique (74.87%), the best bagging (74.42%), and stacking (74.39%) on average. Figure 2 shows the classification accuracy of plain best (PB), bagging of each best plain technique (bagging of decision tree [B-DT], bagging of logistic regression [B-LR], bagging of artificial neural network [B-ANN], bagging of Bayesian network [B-BN], and bagging of Naïve Bayes [B-NB]), stacking, and stacking-bagging method in each dataset

3. Results from Sequential Patterns Analysis

As mentioned in Section 3.4.2, we set the minimum support as 1%, 3%, and 6%, and the minimum confidence as 10%, 15% and 35% in sequence dataset 1, 2, and 3, respectively. Since it is hard to derive sequential patterns with high support from real world data in medical domain and the proper minimum support depends both on the characteristics of problems to be solved and on the policy of the institute which will use the sequential patterns, we set the minimum support to a low value, similar to the one which other researchers have set [26]. Although the minimum support is low, the sequential rules found in this study may carry a significant meaning in preventing foreseeable diseases of revisiting patients. Since not all of the sequential rules are appropriate to predict other possible diseases, selecting useful and meaningful sequential patterns should be conducted carefully. For example, when considering the sequential patterns of length 2, since many patients have cold which is a relatively common disease, it seems that sequential patterns of length 2 including cold are meaningless. So, we did not report them in Table 4. However, when considering the sequential patterns of length more than 2, 'cold' which is associated with other diseases in those sequential patterns may be used to predict foreseeable diseases, as the 9th, 15th, and 16th sequential patterns in Table 4. As shown in the Table 4, total 16 sequential patterns were found. From the identified sequential patterns, we can know that hypertension and bronchitis are frequently associated with other diseases. For instance, other diseases such as hyperlipemia and diabetes mellitus can lead to hypertension, and vice versa. Furthermore, we calculated the average time gap (represented in parenthesize at the end of each sequential pattern in Table 4) between antecedent and consequent diseases in each sequential rule. This time gap can be used to predict when the associated diseases may arise from the antecedent diseases approximately.

IV. Discussion

Much data has been accumulated in many organizations. Al-

Table 4. Sequential patterns found in all sequential dataset

No.	Chain length	Sequential rules
Sequence dataset 1		
1	2	ESSENTIAL HYPERTENSION → HYPERLIPEMIA (152 days)
2	2	NO COMPLICATION NON-INSULIN-DEPENDENT DIABETES MELLITUS → ESSENTIAL HYPERTENSION (65 days)
3	2	HYPERLIPEMIA → ESSENTIAL HYPERTENSION (59 days)
4	2	CHRONIC BRONCHITIS → ESSENTIAL HYPERTENSION (71 days)
5	2	ARTHRITIS → ESSENTIAL HYPERTENSION (126 days)
Sequence dataset 2		
6	2	ESSENTIAL HYPERTENSION → NO COMPLICATION NON-INSULIN-DEPENDENT DIABETES MELLITUS (82 days)
7	2	ARTHRITIS → ARTHRITIS BUNDLE (94 days)
8	2	GASTRITIS → CHRONIC BRONCHITIS (61 days)
9	3	ESSENTIAL HYPERTENSION → COLD → CHRONIC BRONCHITIS (222, 41 days)
Sequence dataset 3		
10	2	CHRONIC BRONCHITIS → GASTRITIS (125 days)
11	2	CHRONIC BRONCHITIS → PERIPHERAL VASCULAR DISEASE (33 days)
12	2	ARTHRITIS SHOULDER → CHRONIC BRONCHITIS (35 days)
13	2	HYPERLIPEMIA → NON-INSULIN-DEPENDENT DIABETES MELLITUS (36 days)
14	3	ARTHRITIS BUNDLE → ARTHRITIS SHOULDER → CHRONIC BRONCHITIS (42, 35 days)
15	3	ARTHRITIS BUNDLE → COLD → CHRONIC BRONCHITIS (10, 32 days)
16	3	COLD → NO COMPLICATION NON-INSULIN-DEPENDENT DIABETES MELLITUS → CHRONIC BRONCHITIS (221, 106 days)

though we have been saying that data is an important asset, they are still not utilized to its maximum extent. Recognizing that most public health centers collect medical records of visiting patients every day without attempting to utilize it, we discussed with the staffs of a public health center in Korea and decided to analyze its data in order to enhance the managerial efficiency of the center and to help the center provide better medical service to its patients.

Through the analysis of the public health center, we aimed to find answers to the following questions: 1) Can we predict whether a patient will revisit the center?; 2) Can we suggest foreseeable disease to the patients who revisit the center? We built 12 different classification models and compared their classification accuracy to find a solution to the first question in each dataset and carried out sequential pattern analysis to provide an answer to the second.

From the results of our classification analysis, we found out these: 1) in general, most influential variables to determine whether a patient of a public healthcare center will revisit it or not are personal burden, insurance bill, period of prescription, age, systolic pressure, name of disease, and postal code; 2) the best plain classification model is dependent on the dataset (*i.e.*, artificial neural network in 3M data set, decision tree in 6M dataset, and logistic regression in 12M dataset); 3) stacking-bagging method outperformed all the best plain techniques, bagging of each best plain technique, and stacking in all 3 datasets except only stacking in 12M dataset. On average, stacking-bagging method also can give better performance (75.31%) than the best plain technique (74.87%), the best bagging (74.42%), and stacking (74.39%).

From the results of our sequential pattern analysis, we were able to derive 16 sequential patterns among the diseases of revisiting patients. Some of the 16 sequential patterns which may not be well known to general practitioners can give them new insights on predicting foreseeable diseases of patients and providing them with adequate precautionary measures.

In sum, classification models and sequential patterns can help public healthcare centers plan and implement healthcare service programs and businesses which are more appro-

 doi: 10.4258/hir.2010.16.2.67

priate to the local residents, and encourage them to revisit public health centers. In addition, central government can allocate budget and manpower more efficiently and effectively according to the healthcare demand of each local residents estimated by the classification models and sequential patterns.

Our study has a few limitations. Firstly, we analyzed data from only one Korean public healthcare center, so the number of instances in our dataset may not be sufficient to make better induction. Secondly, the data used in our study need to be integrated with those from the hospitals which the patients visited after visiting the public healthcare center, so that more diverse analysis can be conducted with the integrated data. Nonetheless, we believe that such experiments as conducted in our study deserve to be paid attention of public healthcare sector where a huge amount of data still remains unused.

Conflict of Interest

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

References

1. Raphael D, Bryant T. The state's role in promoting population health: public health concerns in Canada, USA, UK, and Sweden. Health Policy 2006; 78: 39-55.
2. Kim HY. A positive study on satisfaction in public medical service according to types of community people. J Korean Assoc Gov 2007; 14: 325-347.
3. Yoon KJ. Using DEA to measure the efficiency of local health centers. J Korean Assoc Policy Stud 1996; 5: 80-109.
4. Yoo KR. Measuring the productivity of public health centers: the case of the Jeollabuk provice in Korea. Korean Rev Public Adm 2003; 37: 261-280.
5. Nam CH. Policy development on health administration system in the era of local autonomous government. J Korean Soc Health Educ Promot 1996; 16: 101-126.
6. Chae YM, Kim HS, Tark KC, Park HJ, Ho SH. Analysis of healthcare quality indicator using data mining and decision support system. Expert Syst Appl 2003; 24: 167-172.
7. Han HK, Kim HS, Sohn SY. Sequential association rules for forecasting failure patterns of aircrafts in Korean airforce. Expert Syst Appl 2009; 36: 1129-1133.
8. Hung SY, Yen DC, Wang HY. Applying data mining to telecome churn management. Expert Syst Appl 2006; 31: 515-524.
9. Wu LG, Lee JX, Huang HD, Liu BJ, Horng JT. An expert system to predict protein thermostability using decision tree. Expert Syst Appl 2009; 36: 9007-9014.
10. Yeh IC, Lien CH. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst Appl 2009; 36: 2473-2480.
11. Lee SM, Park RW. Basic concepts and principles of data mining in clinical practice. J Korean Soc Med Inform 2009; 15: 175-189.
12. Silva A, Cortez P, Santos MF, Gomes L, Neves J. Mortality assessment in intensive care units via adverse events using artificial neural networks. Artif Intell Med 2006; 36: 223-234.
13. Choi JP, Han TH, Park RW. A hybrid Bayesian network model for predicting breast cancer prognosis. J Korean Soc Med Inform 2009; 15: 49-57.
14. Diri B, Albayrak S. Visualization and analysis of classifiers performance in multi-class medical data. Expert Syst Appl 2008; 34: 628-634.
15. Phillips-Wren G, Sharkey P, Dy SM. Mining lung cancer patient data to assess healthcare resource utilization. Expert Syst Appl 2008; 35: 1611-1619.
16. Chang CL, Chen CH. Applying decision tree and neural network to increase quality of dermatologic diagnosis. Expert Syst Appl 2009; 36: 4035-4041.
17. Lee WI, Shin BY. Application of neural networks to recognize profitable customers for dental services marketing: a case of dental clinics in Taiwan. Expert Syst Appl 2009; 36: 199-208.
18. Polat K, Kara SI, Guven A, Gunes S. Usage of class dependency based feature selection and fuzzy weighted pre-processing methods on classification of macular disease. Expert Syst Appl 2009; 36: 2584-2591.
19. Ture M, Tokatli F, Omurlu IK. The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data. Expert Syst Appl 2009; 36: 8247-8254.
20. Chang P, Huang FP, Lai ML. The feasibility of using classification and identification techniques to auto-assess the quality of health information on the web. J Korean Soc Med Inform 2009; 15: 247-254.
21. Kang JO, Chung SH, Suh YM. Prediction of hospital charges for the cancer patients with data mining techniques. J Korean Soc Med Inform 2009; 15: 13-23.
22. Exarchos TP, Papaloukas C, Lampros C, Fotiadis DI. Mining sequential patterns for protein fold recognition. J Biomed Inform 2008; 41: 165-179.

23. Chiang JH, Liu HS, Chao SY, Chen CY. Discovering gene–gene relations from sequential sentence patterns in biomedical literature. Expert Syst Appl 2007; 33: 1036-1041.
24. Ryan GW. What do sequential behavioral patterns suggest about the medical decision-making process: modeling home case management of acute illnesses in a rural Cameroonian village. Soc Sci Med 1998; 46: 209-225.
25. Lasker GE. Application of sequential pattern-recognition technique to medical diagnostics. Int J Biomed Comput 1970; 1: 173-186.
26. Lin F, Chou S, Pan S, Chen Y. Mining time dependency patterns in clinical pathways. Int J Med Inform 2001; 62: 11-25.
27. Concaro S, Sacchi L, Bellazzi R. Temporal data mining methods for the analysis of the AHRQ archives. In: Proc Am Med Inform Assoc 2007 Annu Symp. Chicago (IL): American Medical Information Association; 2007.
28. Cooper GS, Virnig B, Klabunde CN, Schussler N, Freeman J, Warren JL. Use of SEER-Medicare data for measuring cancer surgery. Med Care 2002; 40(8 Suppl): IV-43- IV-48.
29. Earle CC, Nattinger AB, Potosky AL, Lang K, Mallick R, Berger M, Warren JL. Identifying cancer relapse using SEER-Medicare data. Med Care 2002; 40(8 Suppl): IV-75-IV-81.
30. Oh HS. Health behavior and utilization of the public health center in rural adults. J Honam Univ Acad 2008; 29: 361-376.
31. Yoo HR. Implementing a smoking cessation clinic at a public health center in Korea: evaluating the outcomes and the smokers' perceptions. J Korean Acad Public Health Nurs 2008; 22: 62-73.
32. Yoon HS, Lee HY, Lee SK. Factors associated with the use of health promotion program: Seoul community health center. Health Soc Welf Rev 2008; 28: 157-184.
33. Cao JM, Kanafani A. Real-time decision support for integration of airline flight cancellations and delays part II: algorithm and computational experiments. Transp Plan Technol 1997; 20: 201-217.
34. Devaraj S, Kohli R. Information technology payoff in the health-care industry: a longitudinal study. J Manag Inf Syst 2000;16:41–67.
35. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 2000; 40: 139-157.
36. Opitz D, Maclin R. Popular ensemble methods: an empirical study. J Artif Intell Res 1999; 11: 169-198.
37. Zhu D. A hybrid approach for efficient ensembles. Decis Support Syst 2010; 48: 480-487.

doi: 10.4258/hir.2010.16.2.67