KIPS Tr. Software and Data Eng.
Vol.6, No.6 pp.303~308 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2017.6.6.303

Health Examination Data Based Medical Treatment Prediction by Using SVM

Minghao Piao† · Jeong-Yong Byun††

ABSTRACT

Nowadays, living standard is improved and people have high interest to the personal health care problem. Accordingly, people desire to know the personal physical condition and the related medical treatment. Thus, there is the necessary of the personalized medical treatment, and there are many studies about the automatic disease diagnosis and the related services. Those studies focus on the particular disease prediction which is based on the related particular data. However, there is no studies about the medical treatment prediction. In our study, national health data based medical treatment predictor is built by using SVM, and the performance is evaluated by comparing with other prediction methods. The experimental results show that the health data based medical treatment prediction resulted in the average accuracy of 80%, and the SVM performs better than other prediction algorithms.

Keywords : Health Examination Data, Medical Treatment Prediction, Data Mining, SVM

SVM을 이용한 건강검진정보 기반 진료과목 예측

Minghao Piao† · 변 정 용††

요 약

생활 수준의 향상 및 소비자들의 건강에 대한 관심의 증가로 인해 자신의 건강에 대해서 스스로 결정하고자 하는 요구가 점차 증가하고 있다. 이로 인해 개인 맞춤형 의료에 대한 요구가 높아지고 있으며 각종 의료 정보를 기반으로 하는 질병 진단에 대한 연구가 많이 진행되고 있다. 하지만 기존의 연구들은 특정 질환과 관련된 데이터를 이용한 특정 질환 예측을 위한 것으로 진료과목을 예측한 연구는 없었다. 본 논문에서는 국민건강정보데이터를 이용하여 진료과목 예측에 관한 연구를 진행하였다. 실험 결과에서 보여주다시피 일반 건강검진 데이터를 이용하여 진료과목을 예측한 결과 평균 80% 이상의 정확도를 보여 주고 있으며 SVM은 다른 예측 알고리즘들보다 뛰어난 성능을 보여 주었다.

키워드 : 건강검진정보, 진료과목 예측, 데이터 마이닝, SVM

1. 서 론

맞춤형 의료 서비스는 개인의 의료정보를 바탕으로 건강에 관한 정보를 제공하고 개인별 맞춤 건강관리 지침을 제공받는 일체 서비스를 말한다. 이러한 서비스를 목적으로 비만, 대사증후군, 뇌졸중 등에 대한 많은 선행연구들이 있었다[1]. 그 외에도 관상동맥 질환 예측[2], 신장 투석 생존율 예측[3], 폐암 예측[4], 간질환 예측[5], 폐암 조기 진단[6], 피부암 조기 발견[7], 심장질환 예측[8-10], 당뇨병 및 간염 진단[10], 유방암 예측[11] 등에 대한 연구들이 있었다. 이외에도 다수의 연구들이 존재하지만 기존의 연구들은 모두 특정 질환을 예측의 대상으로 모델링을 진행하였다. 이러한 연구들은 특정 질환과 관련된 데이터를 이용한 연구들이여서 데이터가 비공개인 경우가 많고 공개 데이터일지라도 다른 연구에는 적용이 불가능하다. 특히 진료과목을 예측하기 위한 연구는 진행되지 않았다. Table 1은 다양한 예측기법들을 이용한 기존의 질환 예측 관련 연구들을 보여 주며 본 논문과의 차이점도 보여준다. 본 연구에서는 다른 연구들과 달리 특정 질환이 아닌 진료과목 예측을 목표로 하였다[12].

국민건강보험공단에서 보유하고 있는 데이터는 다양한 정보들로 이루어져 있다. 이러한 데이터를 이용하여 국민건강보험공단에서는 비만 개선, 건강나이 알아보기, 뇌졸중 예측 프로그램, 대사증후군 맞춤 정보 제공 등 서비스를 제공해왔다. 하지만 데이터의 비공개로 인해 일반 연구자들이 관련

† 정 회 원 : 동국대학교 컴퓨터공학과 연구교수
†† 종신회원 : 동국대학교 컴퓨터공학과 교수
Manuscript Received : October 28, 2016
First Revision : December 30, 2016
Second Revision : February 1, 2017
Accepted : March 6, 2017
* Corresponding Author : Jeong-Yong Byun(byunjy@dongguk.ac.kr)

Table 1. Comparison of Related Studies

Related works	Used methods	Prediction target
Verma [2]	Multi-layer perceptron, Multinomial logistic regression, Fuzzy unordered rule induction algorithm, C4.5	Coronary artery disease
Lakshmi [3]	Artificial neural network, Decision tree, Logical regression	Survival rate of kidney dialysis
Krishnaiah [4]	Rule based, Decision tree, Naïve bayes, Artificial neural network	Lung cancer
Shazmeen [5]	Decision tree, Artificial neural network, Support vector machine, K-nearest neighbor, Naïve bayes	Liver disease
Ahmed [6, 7]	K-means, Frequent pattern mining	Lung cancer risk, Skin cancer risk
Taneja [8]	Decision tree, Naïve bayes, Artificial neural network	Heart disease
Shouman [9]	Single and Hybrid data mining techniques	Heart disease
Kumar [10]	ID3, C4.5, CART	Heart disease, Diabetes, Hepatitis
Piao [11]	Incremental decision tree	Breast cancer
Proposed	SVM	Medical treatment

연구를 진행하는 것은 어려웠다. 정책 3.0에 맞추어 개방한 국민건강정보데이터[13]는 진료과목 예측, 질환 치료 비용 예측 등 다양한 시도를 할 수 있는 정보를 함유하고 있다.

따라서 본 논문에서는 국민건강정보데이터를 활용한 데이터 마이닝 기반의 진료과목 예측을 진행하였다. 예측 모델링을 위한 학습데이터로 {건강검진정보}와 {진료내역정보}를 사용하였으며 서포트 벡터 머신 (SVM)[14]을 사용하여 진료과목 예측 모델을 생성하였다. 2장에서는 국민건강정보데이터에 대한 설명을 하고 3장에서는 서포트 벡터 머신 및 진료과목 예측 모델링 절차에 대해서 설명한다. 4장에서는 모델링에 따른 실험결과를 보여주고 5장에서는 결론과 향후 연구에 대해서 설명한다.

2. 국민건강정보데이터

국민건강정보데이터는 2002년부터 2013년 기간에 해당하는 국민건강보험가입자 100만 명의 {진료내역정보} (Medical Treatment), {의약품처방정보} (Medicine Prescription) 및 {건강검진정보} (Medical Examination) 등 세 부분으로 구성되어 있다. Table 2는 데이터의 구성을 보여준다. {진료내역정보}, {의약품처방정보} 및 {건강검진정보}는 가입자일련번호로 통합이 가능하지만 개인정보 보호를 원칙으로 데이터를 각각 무작위로 추출한 것이므로 모든 진료환자의 데이터가 가입

Table 2. Description of National Health Data

Medical Treatment	Examinee information: sex, age, residence and etc.
	Treatment details: main-sick, sub-sick, convalescence days, hospitalization days, prescription days and etc.
	Medical care expenses claim inspection results.
Medicine Prescription	Examinee information: sex, age, residence and etc.
	Prescription details: prescription ID, convalescence starting day, daily · once dosage, total dosage days, cost and etc.
Medical Examination	Examinee information: sex, age, residence and etc.
	Medical examination result: height, weight, waist, blood pressure, blood sugar, cholesterol, protein in urine, sight, hearing, smoking, drinking and etc.

자일련번호로 연결되지 않는다. 그러므로 연구 목적에 따라서 데이터를 통합할 때 생성된 데이터 크기는 원본 데이터보다 작아지게 된다.

3. 서포트 벡터 머신 및 진료과목 예측 모델링

3.1 서포트 벡터 머신

서포트 벡터 머신(Support Vector Machine, SVM) [14]은 데이터 마이닝 알고리즘의 하나로 지도 학습 모델이며 주로 분류, 예측 등에 사용된다. 주어진 데이터가 두 개의 클래스에 속한다고 가정하였을 때 SVM 알고리즘은 주어진 데이터 집합을 바탕으로 분류 모델을 생성하여 새로운 데이터가 어느 클래스에 속할지 판단하는 비확률적 이진 선형 분류 모델을 만든다. Fig. 1과 같이 만들어진 분류 모델은 데이터가 사상된 공간에서 경계로 표현되는데 SVM 알고리즘은 그 중 가장 큰 폭(margin)을 가진 경계(hyperplane)를 찾는 알고리즘이다. SVM은 선형 분류와 더불어 비선형 분류에서도 사용될 수 있다. 비선형 분류를 하기 위해서 주어진 데이터를 고차원 특징 공간으로 사상하는 작업이 필요한데, 이를 효율적으로 하기 위하여 커널을 사용하며 커널은 선형과 비선형으로 나뉜다.

건강검진 데이터의 경우, 신체정보 및 병리검사결과에 대한 정보 등을 포함하고 있는 다양한 속성들을 가지고 있으며 이러한 속성들은 범주형 또는 연속형 값들을 가지고 있

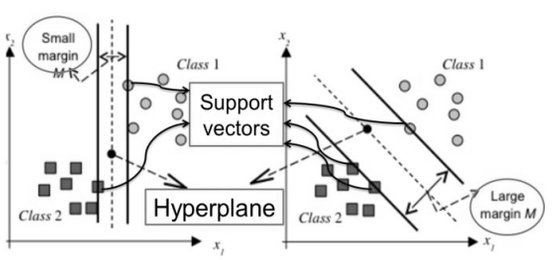

Fig. 1. Support Vector Machine

다. 이는 다차원공간에서 비선형 경계를 가지고 있을 확률이 높으므로 비선형 분류를 하는 것이 적합하다. 따라서 본 연구에서는 비선형 커널인 Polynomial kernel을 사용하여 훈련 데이터를 기반으로 예측 모델을 생성하였다.

3.2 진료과목 예측 모델링

데이터 마이닝을 이용한 예측 모델링은 (1) 데이터 선택, (2) 데이터 전처리 및 변환, (3) 알고리즘 적용을 통한 모델링, 그리고 (4) 평가 단계로 구성된다. (1) 데이터 선택 단계에서는 연구에 필요한 데이터를 선택하는 것을 말한다. 다수의 데이터 집합이 있으면 연구 목적에 맞게 데이터를 선택하고 통합하여야 한다. (2) 데이터 전처리 및 변환 단계에서는 데이터의 질을 보장하기 위하여 선택된 데이터에 대해서 필요한 전처리를 진행한다. 전처리는 주로 데이터의 불완전성(incompleteness), 잡음(noise) 및 모순된 부분(inconsistency)을 제거하는데 초점을 둔다. 예를 들어서 연속형 속성의 결측치는 그 속성이 가지고 있는 값들의 평균값으로 대체할 수 있다. 경우에 따라서는 퓨리에 변환 등 데이터의 형태를 바꾸기 위한 변환을 진행하여야 한다. (3) 알고리즘 적용을 통한 모델링 단계에서는 데이터 특성에 맞는 알고리즘들을 선택하여 비교하고 분석하거나 새로 개발하여 학습시키는 것을 말한다. (4) 평가 단계에서는 생성된 예측 모델의 성능 및 정확도 등에 대해서 비교 및 분석을 진행한다.

Fig. 2는 데이터 마이닝 단계에 따른 국민건강검진정보데이터를 기반으로 진료과목을 예측하기 위한 프레임워크를 보여준다. (1) 데이터 선택: 본 연구에서는 건강검진정보를 이용하여 진료과목을 예측하는 것이므로 건강검진에 대한 정보를 가지고 있는 {건강검진정보} 및 진료과목 등 진료내역에 대한 정보를 가지고 있는 {진료내역정보}를 선택하였다. 선택된 데이터 중 예측 모델링을 위하여 2013년 건강검진정보 및 진료내역정보를 가입자일련번호로 통합을 진행하였다. (2) 데이터 전처리: 결측치가 포함된 데이터는 전체 레코드를 삭제하였으며, 가입자별 연간 최대 방문한 진료과목을 훈련 및 예측의 대상으로 정하였다. (3) 예측 모델링: 전처리된 데이터를 이용하여 훈련 데이터를 생성하였으며 SVM을 적용하여 진료과목 예측을 위한 모델을 구축하였다. (4) 예측 모델 평가: 구축된 예측 모델의 정확도는 독립적인 훈련 및 테스트 데이터들을 이용하여 평가하였고 기타 예측 알고리즘들과의 비교, 분석을 통하여 SVM의 성능을 평가하였다.

4. 실험 결과

건강검진정보 및 진료내역정보를 통합, 정제하여 최종 517개의 데이터를 추출하였으며 속성별 상세 설명은 Table 3과 같다. 진료과목 중 예측의 대상으로 사용된 진료과목은 5개로 내과 (Internal medicine, 204개), 소아청소년과(Pediatrics, 103개), 정형외과 (Orthopedics, 96개), 산부인과(Maternity unit, 59개), 이비인후과(Ear-nose-and-throat department, 55개)이다. 실험을 위하여 사용한 컴퓨터의 사양은 윈도우 10, 메모리 (8GB), CPU (i5, 3.20GHz)이다. 사용한 분석도구는 SPSS사의 Clementine 12.0이다.

4.1 훈련 및 테스트 데이터 생성

Table 3은 건강검진정보 및 진료내역정보를 가입자일련번호로 통합하여 생성한 데이터를 보여준다. 데이터는 모델 생성을 위한 훈련 데이터와 모델 평가를 위한 테스트 데이터로 구분하였다.

모델을 훈련하기 위한 훈련 데이터는 전체 데이터에서 랜덤으로 65%를 여섯 번 선택하였고, 테스트 데이터 또한 랜덤으로 50%를 다섯 번 선택하였다. 6개의 훈련 데이터를 이용하여 서로 다른 예측 모델을 6개 만들었으며, 각 모델별로 다섯 번의 테스트를 진행하였다.

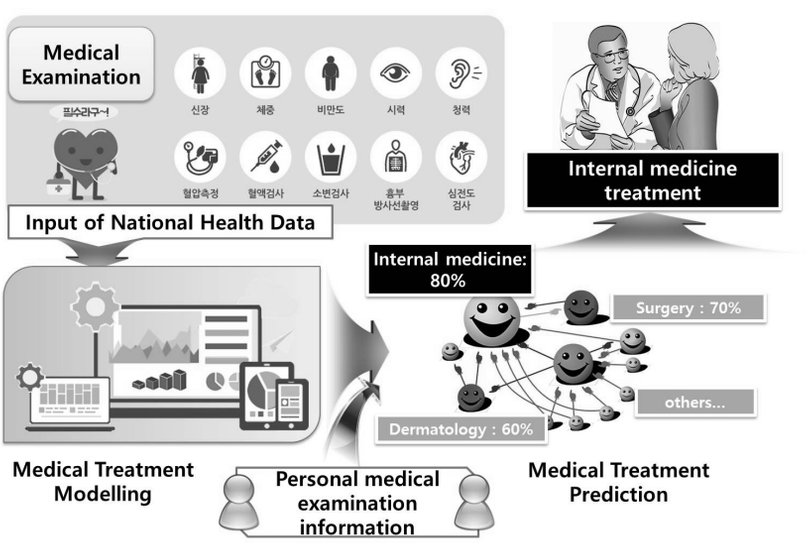

Fig. 2. National Health Data Based Medical Treatment Prediction Framework

Table 3. Details of the Used Data

Item	Check list	Value
Height (cm), Weight (Kg), Waist (cm)		
Hypertension	Maximum blood pressure	mmHg
	Minimum blood pressure	mmHg
Diabetes	Fasting blood glucose	mg/dL
Hypertension, Dyslipidemia, Artery hardening	Total cholesterol	mg/dL
	LDL-cholesterol	mg/dL
	Neutral fat	mg/dL
	HDL-cholesterol	mg/dL
Kidney disease	Protein in urine	1, 2, 3, 4, 5, 6
Anemia	Hemoglobin	g/dL
Chronic kidney failure	Serum creatinine	mg/dL
Liver diseases	AST (SGOT)	U/L
	ALT (SGPT)	U/L
	y-GPT	U/L
Eyesight test	The left	0.1~2.0
	The right	0.1~2.0
Hearing test	The left	1 (normal), 2 (abnormal)
	The right	1 (normal), 2 (abnormal)
Tooth health	Wisdom tooth	1 (no), 2 (yes)
Smoking	1 (does not smoke), 2 (smoked but kicked), 3 (smoking)	
Drinking	1 (does not drinking), 2 (drinking)	
Medical treatment		

4.2 실험 결과 및 분석

Table 4는 상세한 실험 결과를 보여준다. 6개의 SVM 기반 진료과목 예측 모델들은 5개의 테스트 데이터에서 평균 80% 이상의 정확도를 보여주고 있다. 이는 SVM이 진료과목 예측 모델링에 있어서 효율적이라는 것을 증명한다. 특히 내과(inter medicine)와 소아청소년과(pediatrics)에서 높은 정확도를 보여 주었다. 가장 낮은 정확도를 보여준 진료과목은 이비인후과(ear-nose-and-throat department)이며 평균 60~70%의 정확도를 보여 주었다. 내과와 소아청소년과에서 높은 정확도를 보여준 원인은 내과 204개, 소아청소년과 103개의 충분한 훈련 데이터로 인해 예측 모델이 충분히 학습되었기 때문이다. 반대로 산부인과와 이비인후과에서 정확도가 낮은 원인은 각각 59개, 55개의 비교적 적은 훈련 데이터로 인해 충분한 학습이 진행되지 않았기 때문이다. 충분한 데이터를 확보하면 진료과목 예측 모델의 정확도는 더 높아질 것으로 보인다.

Table 5는 SVM과 기타 예측 알고리즘들의 정확도 비교를 보여준다. 비교 분석을 위하여 랜덤으로 테스트 데이터들을 추출하였으며 동일한 테스트 데이터에서 예측 모델들의 정확도를 비교하였다. 다섯 번의 테스트 모두 SVM이 다른 예측 알고리즘들보다 높은 정확도를 보여주었다. 실험에 사용한 데이터는 범주형 및 연속형 속성 값들을 동시에 가지고 있으므로 비선형적인 경계를 가지고 있을 확률이 높다. 따라서 단순히 선형이나 비선형 분류에 뛰어난 알고리즘들보다 비선형 분류를 할 수 있을 뿐만 아니라 비선형 분류 문제를 선형 분류 문제로 바꿀 수 있는 SVM이 더 좋은 성능을 보여주었다.

Table 4. Predicted Results of SVM

Model	Medical treatment	Test-1	Test-2	Test-3	Test-4	Test-5
Model-1	Internal medicine	85.58	87.74	87.5	89.11	81.11
	Maternity unit	80.65	83.33	84.85	76.67	82.76
	Pediatrics	86.54	85.11	86.05	82.46	82.46
	Ear-nose-and-throat department	67.86	75	72	74.19	78.95
	Orthopedics	80.39	79.55	81.81	83.02	80
	Overall	82.33	83.78	84.02	83.46	81.28
Model-2	Internal medicine	85.42	81.55	88.89	84.21	85.26
	Maternity unit	93.75	88.89	75.86	88	82.61
	Pediatrics	88.89	82	88.89	94.34	91.3
	Ear-nose-and-throat department	61.9	75	61.29	53.85	64
	Orthopedics	71.74	75	77.5	70.45	69.77
	Overall	82.73	80.65	82.21	81.07	81.03
Model-3	Internal medicine	82.8	82.56	85.71	78.18	82.46
	Maternity unit	81.48	89.29	86.21	84.85	81.82
	Pediatrics	91.94	92.31	92.98	90.91	92.16
	Ear-nose-and-throat department	57.69	68.18	65.62	66.67	70
	Orthopedics	79.59	83.67	86.36	82.61	81.63
	Overall	81.71	84.39	85.02	80.61	82.67

Table 4. (Continued)

Model	Medical treatment	Test-1	Test-2	Test-3	Test-4	Test-5
Model-4	Internal medicine	84.69	84.76	87.13	88.18	85.42
	Maternity unit	70.97	72.41	72.23	70.97	79.31
	Pediatrics	88.89	91.07	95.08	94.64	90.2
	Ear-nose-and-throat department	63.64	72	78.12	70.97	70.97
	Orthopedics	77.97	71.43	68.18	67.35	71.43
	Overall	80.39	80.81	83.03	81.95	81.25
Model-5	Internal medicine	85.71	89.8	87.74	88.78	86.54
	Maternity unit	78.79	88	76	87.5	76.92
	Pediatrics	90	87.27	85.71	85.48	84.09
	Ear-nose-and-throat department	77.14	80.77	76.92	61.54	81.48
	Orthopedics	66.07	71.43	69.23	64.44	67.44
	Overall	80.77	84.58	82.04	80.78	81.15
Model-6	Internal medicine	89.32	86.46	82.5	86.79	86.81
	Maternity unit	74.07	74.29	64.52	77.42	72.22
	Pediatrics	87.93	90.91	98.25	89.58	94.12
	Ear-nose-and-throat department	62.96	70.37	69.57	75	66.67
	Orthopedics	72.92	75.76	78.57	72.22	74
	Overall	81.75	82.52	82.05	82.13	81.75

Table 5. Comparison of SVM with Other Prediction Methods

Algorithm	Test-1	Test-2	Test-3	Test-4	Test-5
SVM	80.86	80.3	80.52	80.72	84.25
C4.5	75.76	73.23	74.53	72.29	75.59
ANN	56.06	51.67	56.55	51.81	53.54
Bayes Net	70.45	66.54	68.91	68.67	66.54
Logistic	65.15	63.94	63.3	62.25	63.39

5. 결 론

본 논문에서는 SVM을 이용한 건강검진정보데이터 기반의 진료과목에 대한 예측 모델링을 진행하였다. 실험 결과에서 보여주다시피 건강검진정보를 기반으로 진료과목에 대해 예측한 결과 평균 80% 이상의 정확도를 보여 주었다. 또한 SVM은 다른 알고리즘들보다 뛰어난 결과를 보여 주었다. 일부 진료과목은 훈련 데이터가 충분하지 않은 관계로 인해 다른 진료과목에 비해서 예측이 제대로 진행되지 않았지만 충분한 데이터를 확보하면 더욱 정확한 모델을 구축하여 예측의 정확도를 높일 수 있을 것으로 보인다.

본 논문의 경우 확보한 데이터의 한계로 인하여 모델링을 위한 데이터를 많이 추출하지 못하였다. 또한 진료과목별 데이터 분포가 일정하지 못하여 일부 진료과목들은 예측의 대상에서 제외되었다. 따라서 향후 정확한 통합이 가능한 국민건강정보데이터를 확보하여 더욱 많은 진료과목을 예측할 수 있는 모델을 구축하고 다양한 연구를 진행할 예정이다.

References

[1] G. H. Cho, Y. M. Park, S. H. Ji, J. E. Choo, and H. S. Im, “Development of personalized-integrated health care program,” Research report, National Health Insurance Service Ilsan Hospital, No.2014-20-010, 2014.

[2] L. Verma, S. Srivastava, and P. C. Negi, “A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data,” *Journal of Medical Systems*, Vol.40, pp.1-7, 2016.

[3] K. R. Lakshmi, Y. Nagesh, and M. VeeraKrishna, “Performance comparison of three data mining techniques for predicting kidney disease survivability,” *International Journal of Advances in Engineering & Technology*, Vol.7, Issue.1, pp.242-254, March 2014.

[4] V. Krishnaiah, “Diagnosis of lung cancer prediction system using data mining classification techniques,” *International Journal of Computer Science and Information Technologies* (IJCSIT), Vol.4, No.1, pp.39-45, 2013.

[5] S. F. Shazmeen, M. M. A. Baig, and M. R. Pawar, “Performance evaluation of different data mining classification algorithm and predictive analysis,” *Journal of Computer Engineering*, Vol.10, No.6, pp.1-6, 2013.

[6] K. Ahmed, A. A. Emran, T. Jesmin, R. F. Mukti, M. Z. Rahman, and F. Ahmed, “Early detection of lung cancer risk using data mining,” *Asian Pacific Journal of Cancer Prevention*, Vol.14, pp.595-598, 2013.

[7] K. Ahmed, T. Jesmin, and M. Z. Rahman, “Early prevention and detection of skin cancer risk using data mining,” *International Journal of Computer*, Vol.62, No.4, pp.1-6, 2013.

[8] A. Taneja, “Heart disease prediction system using data mining techniques,” *Oriental journal of Computer science & technology*, Vol.6, No.4, pp.457-466, December 2013.

[9] M. Shouman, T. Turner, and R. Stocker, “Using data mining techniques in heart disease diagnosis and treatment,” in *Proceedings of Japan-Egypt Conference on Electronics, Communications and Computers*, IEEE, Vol.2, pp.174-177, 2012.

[10] D. S. Kumar, G. Sathyadevi, and S. Sivanesh, “Decision support system for medical diagnosis using data mining,” *Journal of Computer Science*, Vol.8, No.3, pp.147-153, 2011.

[11] M. H. Piao, J. B. Lee, K. E. Saeed, and K. H. Ryu, “Discovery of significant classification rules from incrementally inducted decision tree ensemble for diagnosis of disease,” in *Proceedings of International Conference on Advanced Data Mining and Applications*, pp.587-594, 2009.

[12] Shinyoung Ahn, Yookyung Lee, Minghao Piao, and Jeongyong Byun, “National Health Data based Medical Treatment Prediction,” *2016 Fall KIPS Conference*, Vol.23, No.2, pp.546-547, 2016.

[13] Naional Health Insurance Service, “National Health Data User Manual [ver 1.0],” 2016.

[14] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, Vol.20, No.3, pp.273-297, 1995.

Minghao Piao

e-mail : myunghopark@gmail.com

2007년 중국 연변과학기술대학교 컴퓨터과학과 기술학과(학사)

2009년 충북대학교 바이오인포메틱스학과 (석사)

2014년 충북대학교 컴퓨터과학과(박사)

2015년~현 재 동국대학교 컴퓨터공학과 연구교수

관심분야 : Data Mining, Big Data, Bioinformatics

변 정 용

e-mail : byunjy@dongguk.ac.kr

1980년 동국대학교 컴퓨터공학과(학사)

1983년 동국대학교 컴퓨터공학과(석사)

1994년 홍익대학교 컴퓨터공학과(박사)

1988년~현 재 동국대학교 컴퓨터공학과 교수

관심분야 : Database Management Systems, Semantic Web and Web Service, Korean Alphabet