Journal of Korea Multimedia Society Vol. 16, No. 3, March 2013(pp. 269-280)
http://dx.doi.org/10.9717/kmms.2013.16.3.269

망막 질환 진단을 위한 베이지안 네트워크에 기초한 데이터 분석

김현미[†], 정성환[††]

요 약

본 논문에서 망막 질환 요인간의 의존도 분석을 위해 효율적인 분류기를 활용할 수 있는 방안을 제시하였다. 먼저 여러 베이지안 네트워크 중에서 TAN (Tree-Augmented Naive Bayesian Network), GBN(General Bayesian Network)과 Markov Blanket으로 특징축소된 GBN과의 분류성능과 예측정확률을 비교분석하였다. 그리고 처음으로, 높은 성능을 보인 TAN을 망막 질환 임상데이터의 의존도 분석에 적용하였다. 의존도 분석 결과, 망막 질환의 진단과 예후 예측에 활용의 가능성을 보였다.

Bayesian Network-based Data Analysis for Diagnosing Retinal Disease

Hyun-Mi Kim[†], Sung-Hwan Jung[††]

ABSTRACT

In this paper, we suggested the possibility of using an efficient classifier for the dependency analysis of retinal disease. First, we analyzed the classification performance and the prediction accuracy of GBN (General Bayesian Network), GBN with reduced features by Markov Blanket and TAN (Tree-Augmented Naive Bayesian Network) among the various bayesian networks. And then, for the first time, we applied TAN showing high performance to the dependency analysis of the clinical data of retinal disease. As a result of this analysis, it showed applicability in the diagnosis and the prediction of prognosis of retinal disease.

Key words: Retinal Disease(망막 질환), GBN(일반 베이지안 네트워크), Markov Blanket(마코프 블랭킷), TAN(트리확장 순수 베이지안 네트워크), Prediction of Prognosis(예후 예측)

1. 서 론

임상정보를 포함하는 의료 데이터는 데이터 자체로도 가치가 매우 크다. 이러한 의료 데이터에 대한 분석과 결과를 미래에 유용한 의사결정에 이용하기 위해, 의료 데이터마이닝에 대한 연구가 활발히 이루어지고 있다. 의료 데이터마이닝에서는 베이지안 네트워크(BN: Bayesian Network)가 일반적으로 적용되어, 질병의 진단이나 예측문제에 사용되며 좋은 성능을 보여왔다[1-3].

베이지안 네트워크의 여러 유형 중에서 Friedman 등[4]에 의해 소개된 TAN(Tree-Augmented Naive Bayesian Network)은 NBN(Naive Bayesian Network)의 노드 독립성 가정을 완화하기 위해, 자식노드들 사이에 트리형태의 관계가 있음을 가정한 네트워크이다. 기존의 연구들에서 Jiang 등[5]이 TAN이

※ 교신저자(Corresponding Author) : 정성환, 주소 : 창원시 의창구 사림동 9번지 창원대학교 55호관 55323호실 (641-773), 전화 : 055) 213-3815, FAX : 055) 286-7429, E-mail : sjung@changwon.ac.kr
접수일 : 2012년 3월 20일, 수정일 : 2012년 6월 7일
완료일 : 2012년 12월 24일

[†] 준회원, 창원대학교 컴퓨터공학과
(E-mail : hmkim@changwon.ac.kr)
[††] 종신회원, 창원대학교 컴퓨터공학과
※ 본 논문은 2012 WISET 경남지역사업단의 지원으로 운영되는 '경남미래여성인력양성' 프로그램의 연구결과임.

분류정확성이나 에러비율에서 NBN보다 우수하다고 하였다. 그리고 Chinnasamy 등[6]은 단백질의 구조와 클래스를 예측하기 위해서 TAN을 적용하였다.

또한 Markov Blanket 방법은 주어진 의사결정 문제를 구성하는 변수들 중에서 의미있는 최소한의 변수만을 추출하는 특징선택의 기법으로서 GBN(General Bayesian Network)에서 주로 사용되고 있다. GBN은 베이지안 네트워크 중에서 가장 일반화된 형태이며, 클래스 노드조차 일반 속성노드와 차이를 두지 않고 모든 노드들 간의 상호의존성을 하나의 베이지안 망으로 표현한다. Cheng 등[7]은 GBN에서 Markov Blanket으로 특징을 추출하는 방법이 자료 안에 숨겨진 인과관계를 도출하는데 매우 유용한 방법이라고 하였다. Tsamardinos 등[8]은 Markov Blanket에 속한 특징들에 대한 지식만 가지고 있으면, 클래스 노드의 확률분포를 결정하기에 충분하다고 주장하였다.

본 논문의 연구대상인 망막은 우리 눈의 내부에 있는 얇은 신경막으로 카메라에 비유하면 필름에 해당되는데 한번 손상이 되면 다시는 회복하지 못하는 눈의 중요한 일부이다. 따라서 망막 질환의 진단과 예측을 위해 관련 정보를 제공할 수 있는 연구가 필요하다[9]. 현재까지는 망막 질환 데이터 영역에서는 기계학습이나 베이지안 네트워크 적용이 없었고, 주로 SPSS와 같은 통계툴을 이용하여 요소들 간의 분산분석, 상관계수와 클래스별 분포상태 등을 알아보는 수준이었다[10-12]. 따라서 본 논문에서는 효율적인 분류기를 선택하고 망막 질환 요인들 간의 상호의존도를 분석하기 위해, 망막 임상데이터를 대상으로 베이지안 네트워크를 적용하여 데이터 분석을 처음으로 시도한다.

본 연구의 실험에서는 TAN, GBN과 Markov Blanket 방법으로 특징축소된 GBN 등의 다양한 베이지안 네트워크 기법을 적용하여, 분류성능과 예측정확률을 비교분석한다. 그리고 높은 성능을 보인 분류기를 선택하여, 망막 질환 요인들의 의존도 관계를 분석해 본다. 그리고 실험 결과를 바탕으로 망막 질환의 진단과 예후 예측에 정보 제공의 가능성을 제시하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서 본 연구 주제와 관련한 기존 연구들을 살펴본다. 3장에서는 베이지안 네트워크의 개념과 특징을 살펴보고 TAN, GBN과 Markov Blanket이 적용된 GBN 등의 베이지안 네트워크 분류기를 소개한다. 4장에서는 망막 임상데이터의 의존도 분석을 위한 실험 방법과 결과를 설명한다. 그리고 5장에서 결론 및 향후 연구를 제시한다.

2. 기존 연구

2.1 망막 질환에 적용된 기존 연구 방법

지금까지 보고된 기존 연구에서는 망막 질환 임상데이터에 대하여 주로 SPSS 통계처리를 통한 분석방법을 사용하였다. 먼저 전문가의 진단으로 측정 기준인 IS/OS (junction of Inner Segment and Outer Segment of photoreceptor layer, 망막광수용체층의 경계선)의 상태에 따라 IS/OS 라인이 끊어짐 없이 완전히 관찰되는 정상군, 불연속적으로 존재하는 부분군, 라인이 손상되어 완전히 소멸되어 있는 손상군의 세 그룹으로 분류한다. 연구결과를 살펴보면 Oishi 등[10]과 Michael 등[11]은 IS/OS 등급은 시력(VA: Visual Acuity)과 중심망막두께(CFT: Central Foveal Thickness)와 관련하여 통계적으로 유의한 의미관계가 있다고 주장하였다.

표 1은 본 연구에서 수집한 임상데이터를 가지고 실제로 기존 연구 방법인 SPSS 통계툴로 실험한 결과이다. 상관계수(r)의 값이 $0.3 < |r| < 0.7$ 이면 뚜렷한 상관관계를 의미한다. IS/OS와 중심망막두께는 -0.402이고 IS/OS 클래스와 LogMAR[1] 시력은 +0.645의 측정치가 나왔다. 여기서 IS/OS 클래스는 중심망막두께보다는 LogMAR 시력에서 양의 상관관계로 더욱 유의한 관계가 있음을 알 수 있다.

실제 SPSS 통계툴을 적용한 표 1과 같은 실험결과로 IS/OS 라인의 변성이 심할수록 중심망막두께도 얇아지게 되고 결국 시력도 떨어진다는 것을 알 수 있었다. 그러므로 망막 질환을 가진 환자를 모니터링 할 때 IS/OS 상태가 어느 요소보다도 더 중요한 설계요소가 되어야 한다. 따라서, 기계학습을 통한 훈련으로 자동분류를 가능하게 함으로써 IS/OS 클래스별 분포상태가 자동으로 파악된다면 치료효과를 신속하게 예측할 수 있는 장점이 있기 때문에 이

1) LogMAR 시력은 대수시력표에서의 시력을 말함. LogMAR 시력=-log(시력검사표 시력)

표 1. IS/OS 클래스와 각 속성간의 상관관계

		Age	CFT	LogMAR	IS/OS
Age	Pearson상관계수	1	.149*	.286*	.122
	유의확률(양쪽)		.032	.000	.080
	N	208	208	208	208
CFT	Pearson상관계수	.149*	1	-.182**	-.402**
	유의확률(양쪽)	.032		.009	.000
	N	208	208	208	208
LogMAR	Pearson상관계수	.286**	-.182**	1	.645**
	유의확률(양쪽)	.000	.009		.000
	N	208	208	208	208
IS/OS	Pearson상관계수	.122	-.402**	.645**	1
	유의확률(양쪽)	.080	.000	.000	
	N	208	208	208	208

*상관계수는 0.05수준(양쪽)에서 유의, **상관계수는 0.01수준(양쪽)에서 유의)

에 대한 연구가 필요하다[13].

2.2 베이지안 네트워크가 적용된 기존 연구

베이지안 네트워크 기법이 의학 분야에 적용된 국내의 연구는 다음과 같다. 손호선 등[1]은 심장 질환 데이터를 대상으로 순수 베이지안 분류기와 베이지안 네트워크 분류기로 분류 성능을 비교하였다. 정용규 등[2]은 NBN, GBN, BAN (Bayesian Network Augmented Naive Bayesian) 등의 베이지안 네트워크 분류기들을 불임환자 임상데이터의 분석에 적용하였다. 이제영 등[3]은 정신장애 질병인 섬망(delirium) 분석에 베이지안 네트워크를 사용하여 섬망과 관련된 위험인자들 간의 상호작용을 규명하고 오즈비(odds ratio)를 구하였다.

한편 대표적인 국외 연구로서, Cheng 등[7]은 자료 안에 숨겨진 인과관계를 도출하기 위해 GBN에서 Markov Blanket을 추출하는 방법을 사용하였다. Markov Blanket을 정확하게 구할 수만 있으면 변수들 간의 인과관계를 통하여 클래스 노드에 대한 분류 작업을 정확하게 수행할 수가 있다. 그리고 Tsamardinos 등[8]은 결과변수인 클래스 노드의 확률분포를 결정하기 위해 Markov Blanket에 속한 변수들에 대한 지식만 가지고 있으면 자료 분석에서 충분하다고 하였다. 또한 Chinnasamy 등[6]은 단백질의 구조와 클래스를 예측하기 위해 TAN의 사용을 제안하였다. 기존에 통계적 기술들에 바탕을 둔 판별적(discriminative) 방법들은 분류 클래스 수가 증가함에 따라 예측정확도가 떨어질 수 있다. 그들은 이를 해결하기 위해 TAN을 사용하였으며 클래스 예측성능을 높였다. Jiang 등[5]은 분류정확성이나 에러비율 면에서 TAN이 NBN보다 성능이 우수하다고 하였다.

본 연구에서는 망막 임상데이터 분석에서 효율적인 분류기를 선택하기 위해, TAN이 GBN이나 특징 축소 방법으로 Markov Blanket된 GBN보다 우수한 분류 성능을 보여줄 수 있는지를 실험을 통해 알아보고자 한다.

3. 베이지안 네트워크

3.1 기본 개념

베이지안 네트워크는 확률 값이 모인 집합의 결합확률분포의 결정모델이다. 베이지안 네트워크는 특정분야의 영역지식을 확률적으로 표현하는 대표적인 수단으로, 변수들 간의 확률적 의존관계를 나타내는 그래프와 각 변수별 조건부 확률로 구성된다[14]. 따라서 하나의 BN은 각 노드마다 하나의 조건부확률표(CPT: Conditional Probability Table)를 갖는 비순환유향그래프(DAG: Directed Acyclic Graph)로 정의할 수 있다. 노드와 노드를 연결하는 호(arc)는 노드 사이의 인과관계를 나타내며, 변수의 확률적인 인과관계로 네트워크를 구성하고 조건부확률표(CPT)를 가지고 다음의 식 (1)과 같은 베이지안 정리(Bayesian theorem)을 이용하여 결과를 추론할 수 있다.

$$P(A|B) = \frac{P(A, B)}{P(B)} \quad (1)$$

베이지안 네트워크가 조건부 독립이라 가정하고 곱셈 규칙(product rule)을 적용하면, 네트워크를 구성하는 각 노드에 대한 결합확률은 식 (2)와 같다.

$$P(A,B)=P(A|B)P(B)=P(B|A)P(A) \quad (2)$$

일반적인 베이지안 네트워크는 이와 같은 베이지안 정리, 곱셈 규칙, 체인 규칙(chain rule) 에 의하여 다음과 같은 식 (3)이 만들어진다. 여기서 $x_1,x_2,...,x_n$은 특정 데이터의 속성 집합이고 $Parent(x_i)$는 x_i의 부모 노드들의 집합이다.

$$P(x_1,x_2,...,x_n)=\prod_{i=1}^{n}P(x_i|Parent(x_i)) \quad (3)$$

베이지안 네트워크는 가정의 단순함에도 불구하고 많은 연구를 통해 비교적 높은 분류 성능을 보인다. 베이지안 네트워크는 그림 1과 같이 각 변수들에 대한 의존도를 그래프로 표현할 수 있다. 그림에서 각 변수는 이진 값을 갖는 것으로 가정한다. 심장병(HD)의 부모 노드들은 운동(E)과 식사습관(D)과 같이 병에 영향을 주는 위험 요소에 해당한다. 심장병의 자식 노드들은 고혈압(BP), 흉통(CP)과 같이 심장병의 증상에 해당한다. 예를 들어 가슴앓이(HB)는 건강하지 않은 식사습관으로부터 유발되며 흉통으로 이어질 수 있다. 또한 조건부 확률 값도 가지므로 여기서 알고 있는 확률을 바탕으로 가슴앓이가 있을 때, 고혈압이 있을 때 심장병에 걸릴 확률을 구할 수가 있다[15].

3.2 베이지안 네트워크의 특징

본 연구의 망막 질환 요인들 간의 상호의존도 분석에 사용된 베이지안 네트워크의 장점은 신경망의 분류기와 비교해서 도메인 지식을 적용하기 쉬우며 결과의 분석이 가능하다. 단점으로는 입력값으로 수치값(numeric value) 이 아닌 범위가 정해진 범주값(categorical value)을 사용함으로써 정확도면에서 문제가 생길 수 있으며, 노드수가 많아지면 실험시간이 오래 걸린다. 하지만 의학지식을 적용하여 분석 등이 가능한 의학 도메인에서 도메인 지식 가능성이나 원인분석이 가능하다는 큰 특성이 있다[1]. 현재 데이터 분석과정에 주로 쓰이는 것은 로지스틱 회귀분석, 의사결정트리, 신경망 등의 기법들이다. 표 2는 이러한 데이터 분석기법들과 본 실험에서 적용되는 베이지안 네트워크와의 특징들을 비교한 것이다.

3.3 실험을 위한 베이지안 네트워크 유형들

3.3.1 트리확장 순수 베이지안 네트워크(TAN)

대표적인 유형의 베이지안 네트워크 분류기에는

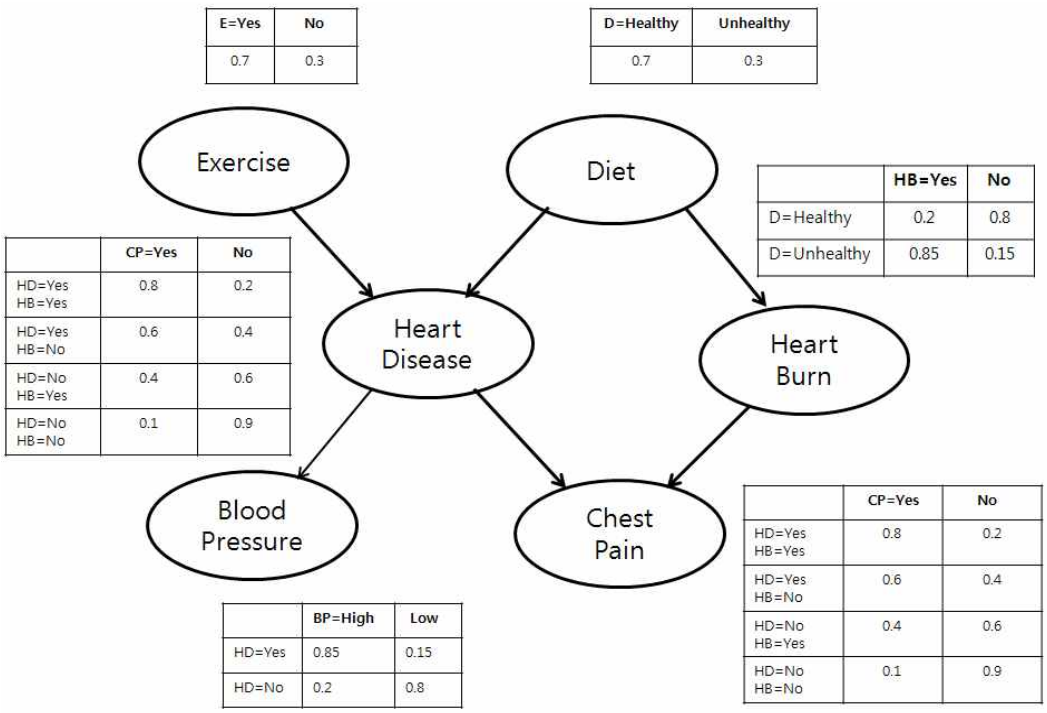

그림 1. 심장병과 위험인자의 관계를 표현한 베이지안 네트워크

표 2. 데이터마이닝 기법들의 장단점 비교

기 법	장 점	단 점
신경망	①분류문제뿐만 아니라 예측, 평가, 합성, 제어 등의 다양한 분야에 적용 ②학습능력을 갖추고 일반화능력이 뛰어나고 구현이 쉬움 ③다층 퍼셉트론은 선형분리가 불가능한 경우에도 높은 성능을 보여주는 한 단계 진보한 신경망	①샘플이 어떤 부류로 분류되었을 때 왜 그런 결정이 내려졌는지 이유를 분석하기 어려움(블랙박스) ②입력값으로 수치형이 아닌 범주형을 사용
로지스틱 회귀 분석	①의학, 보건학 계열에서 다변량 분석으로 많이 쓰임 ②다변량 변수를 독립변수로 하여 종속변수에 미치는 영향을 파악가능 ③입력값으로 수치형과 범주형 모두 취급 가능	①분석자료에 가장 적합한 모델을 선정하는데 시간투자가 필요 ②샘플이 어떤 부류로 분류되었을 때 왜 그런 결정이 내려졌는지 해석하기 어려움(블랙박스)
의사 결정 트리	①의사결정규칙을 도표화하여 관심대상에 해당하는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 계량적 분석 방법 ②샘플이 어떤 부류로 분류되었을 때 왜 그런 결정이 내려졌는지 해석가능(화이트 박스) ③입력값으로 수치형과 범주형 모두 취급 가능	①목표변수가 수치형인 회귀모형에서는 그 예측력이 떨어짐 ②나무가 너무 깊은 경우에는 예측력 저하와 해석이 쉽지 않음 ③가지가 많을 경우 새로운 자료에 적용할 때 예측오차가 큼
베이지안 네트워크	①특정 분야의 영역 지식을 확률적으로 표현하는 대표적인 수단 ②변수들 간의 확률적 의존 관계를 나타내는 그래프와 각 변수별 조건부 확률로 구성 ③분류 클래스 노드의 사후 확률분포를 구해줌으로써 개체들에 대한 하나의 자동분류기로 이용가능 ④샘플이 어떤 부류로 분류되었을때 왜 그런 결정이 내려졌는지 해석가능(화이트박스)	①입력값으로 수치형이 아닌 범주형을 사용 ②노드수가 방대해지면 시간이 오래 소요될 수 있음

그림 2에서와 같이 순수 베이지안 네트워크(NBN), 일반 베이지안 네트워크(GBN)와 트리확장 순수 베이지안 네트워크 (TAN) 등이 있다[16]. NBN은 가정의 단순함에도 불구하고 비교적 높은 분류 성능을 보여주는 것으로 알려져 있다. 그러나 NBN은 클래스노드를 여타 다른 노드와 다른 특별한 변수로 간주하고 변수들 간에 독립성을 지나치게 가정하고 있으므로 현실세계의 현상을 반영하는데 적합하지 않다는 문제점이 제기되었다.

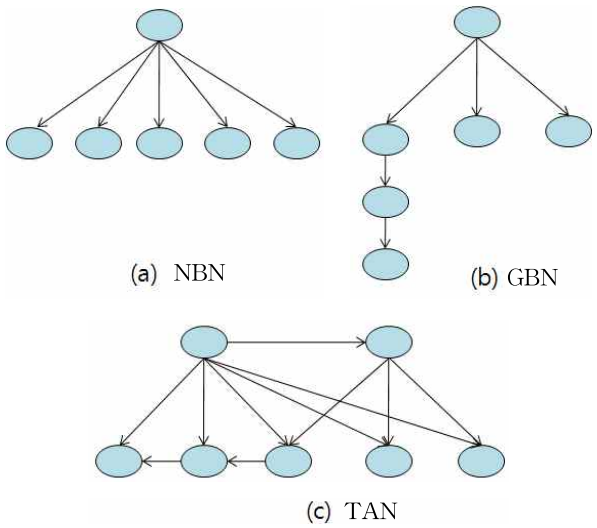

그림 2. 실험에 사용되는 BN 유형들

TAN은 NBN과는 달리 속성 노드들 간에도 상호의존도가 존재한다고 가정하고 이러한 속성 간 상호의존도를 하나의 일반 베이지안 네트워크 형태로 표현 가능하도록 NBN을 확장한 것이다. 즉, TAN은 그림 2와 같이 NBN을 기반으로 트리구조의 네트워크를 결합하는 형태이다. 이는 속성들 간의 의존성을 추가함으로써 나이브 베이지안 분류기의 학습능력을 향상시켜 분류성능을 높인다.

TAN은 식 (4)와 같이 정의된다. 식에서 $A_1,...,A_n$은 속성들의 집합이고 클래스 C의 확률은 P(C)이며 α은 정규화 상수(normalization constant)이다. $parents(A_i)$는 A_i의 부모 노드들의 집합이다.

$$P(C|A,...,A) = \alpha P(C)\prod_{i=1}^{n} P(A_i|parents(A_i)) \qquad (4)$$

3.3.2 GBN과 Markov Blanket으로 특징축소된 GBN

베이지안 네트워크 분류기 중 가장 일반화된 형태는 GBN으로서, GBN에서는 기존 다른 베이지안 네트워크 분류기들과는 달리 클래스 노드조차 일반 특징노드와 차이를 두지 않고 모든 노드들 간의 상호의존도를 하나의 베이지안 네트워크로 표현한 것이다[16]. 따라서 GBN에서는 클래스노드도 부모노드들을 가질 수 있다. Markov Blanket 방법은 주어진 종속 변수를 구분하는데 도움이 되지 않는 설명 변수는 포함시키지 않으며 종속변수에 대한 최소한의 설명변수들로 이루어진다. Markov Blanket은 클래스 노드의 부모노드들과 자식노드들, 그리고 자식노드들의 또 다른 부모노드들을 포함하는 모든 노드들의 부분 집합이다. 4장에서는 Markov Blanket으로 특징 축소된 GBN이 과연 망막 임상데이터 분석에서 신뢰할 만한 결과를 보이는지를 GBN과 TAN 등의 여러 베이지안 네트워크들과 실험해 본다.

4. 실험 및 결과

4.1 실험 데이터

4.1.1 실험 데이터

본 연구에서는 A대학병원 안과학교실의 협조로 104명의 망막 임상데이터를 가지고 우안, 좌안 데이터를 각각 독립적으로 보아 총 208안의 샘플을 대상으로 하였다. 그리고 망막의 손상된 상태에 따라서 정상군(0), 부분군(1), 손상군(2)의 세 그룹으로 분리한다. 데이터 속성들은 표 3과 같이 나이(age), 가족력(fhx), 유전모델(model), 발생시기(onset), 망막두께(ft), OCT상에서 보이는 소견(finding), 황반부종(cme), 황반부종길이(length), 시야검사(gvf), LogMar 시력(va), 분류클래스(isos) 등 총 11가지이다.

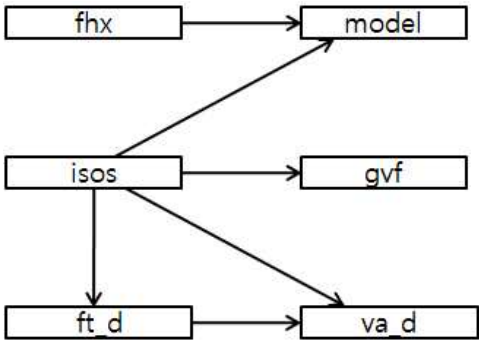

그림 3. BNPC에서 Markov blanket 적용한 GBN

4.1.2 Markov blanket 적용한 GBN의 실험 데이터

GBN의 Markov Blanket으로 이루어진 속성 추출을 위하여, 먼저 Cheng이 구현한 조건부 독립성 검사 기반의 BN PowerConstructor(BNPC)를 이용하여 GBN을 학습한다. 생성된 GBN에서 Markov blanket으로 추출된 속성들은 그림 3과 같다. 속성 집합은 최상위 클래스인 isos(분류 클래스)와 ft_d(망막두께), va_d(시력), gvf(시야검사), model(유전모델), fhx(가족력유무) 등의 총 6가지이다. BNPC에서 실험을 위해서는 속성 값들이 범주형으로 전처리되어야 하는데, ft_d와 va_d는 ft(망막두께)와 va(시력)이 범주화된 형태이다.

표 3. 실험 데이터의 속성 집합(*범주화 필요 속성)

속성	속성 이름	설 명
age	나이	*수치값 (범주화 필요)
fhx	가족력	가족력 유무(0,1)
model	유전 모델	AD(우성), AR(열성), S(독립), X(성관련)
onset	야맹증 발생시기	*수치값 (범주화 필요)
ft	망막 중심 두께	*수치값 (범주화 필요)
finding	OCT상 소견	CME(황반부종), ERM(망막전막), CME+ERM
cme	황반 부종	부종 유무(0,1)
length	황반 부종 길이	0, 1, 5, 10
gvf	시야 검사	진행(0-10), 정상(60-70)
va	시력	*수치값 (범주화 필요)
isos	분류 클래스	정상(0), 부분(1), 손상(2)

4.2 전처리 작업

오픈 데이터마이닝 툴인 Weka를 가지고 베이지안 네트워크를 적용하기 위해서는 데이터 속성 값들이 범주형 값을 가져야한다[17]. 망막 데이터 속성들 중에서 수치형 값인 age(나이), oneset(발생시기), ft(망막두께), va(시력) 등을 범주형 값으로 변환해야 한다. 범주화되는 각 속성들의 범위를 알기 위해 표 4와 같이 BNPC를 통해서 나온 산출물을 참조하였다. 그리고 C++ 로 프로그래밍하여 해당 속성 값들을 범주형으로 변환하였다. 그림 4는 모든 속성들을 범주형으로 변환한 다음에 Weka로 전처리 작업을 하여 나온 산출물이다.

그림 4에서 그림 하단의 isos 영역을 살펴보면 진한색 막대 109명은 정상군(0), 중간색 막대 24명은 부분군(1), 엷은색 막대 74명은 손상군(0)을 나타낸다. 또한 va 영역에서 막대는 차례로 EYE-1, EYE-2, EYE-3, EYE-4를 나타내며, EYE-1은 71명, EYE-2는 56명, EYE-3는 56명, EYE-4는 24명을 나타낸다. EYE-1에서 EYE-4로 갈수록 시력이 나빠지며 EYE-1에서는 진한색의 정상군이 넓게 분포함을 알 수 있다.

표 4. 범주화한 데이터 속성

Attribute	Description	Value	Attribute	Description	Value
age (나이)	<= 30.5 <= 42.5 <= 56 > 56	LOW MID HI HHI	onset (발생시기)	<= 30.5 <= 42.5 <= 56 > 56	LOW MID HI HHI
ft (망막두께)	<= 185.5 <= 219.5 <= 254.5 <= 289 > 289	ONE TWO THREE FOUR FIVE	va (시력)	<=0.185 <= 0.61 <= 2.2 > 2.2	EYE-1 EYE-2 EYE-3 EYE-4

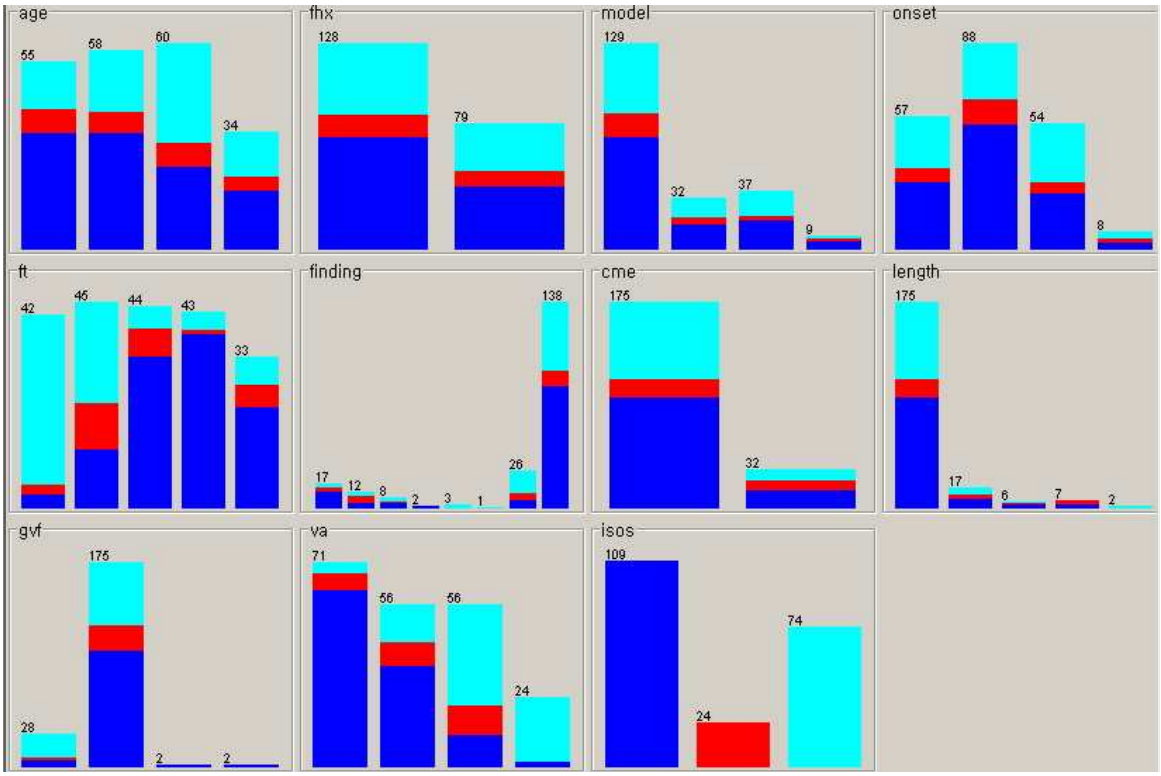

그림 4. Weka에서의 전처리 결과

4.3 실험 결과

Weka를 활용하여 기계학습으로 얻어진 지식을 바탕으로 분류실험을 한 결과는 표 5와 같다. 테스트 집합보다 훈련집합에서의 분류정확률이 높게 나타났고, 10-겹 교차검증의 분류정확률은 훈련집합보다 낮으나 테스트집합보다는 약간 높은 성능을 보여주었다.

TAN, GBN과 Markov Blanket 적용된 GBN의 실험결과를 보면, 훈련집합에서는 크기가 n=30 일 때 Markov Blanket으로 특징 축소된 GBN의 예측정확률이 높지 않았으나 60, 90으로 갈수록 GBN보다 높았다. 또한 테스트 집합과 10-fold 검증집합에서도 GBN보다 높게 나타났다. 그러나 TAN과의 비교에서는 훈련집합, 테스트집합과 10-fold 검증집합에서 모두 Markov Blanket된 GBN이 낮은 성능을 보였다. 이는 Markov Blanket 방법이 최소한의 설명 노드들로 구성하여 복잡도를 줄여 일반화시키는 과정에서 TAN보다 예측정확률이 떨어진다고 판단된다.

표 6과 표 7은 TAN과 Markov Blanket된 GBN의 예측정확성 측면에서 성능을 평가하기 위하여 여러 측정값을 나타내었다. 평균자승오차제곱근(RMSE)은 오차 제곱의 평균에 제곱근을 취한 것으로 표준편차의 정의와 동일하다. TAN에서의 절대평균오차(MAE)은 0.14이며 평균자승오차제곱근(RMSE)은 0.23로 Markov Blanket된 GBN의 MAE 0.28, RMSE 0.37보다 더 낮게 나타났다. 따라서 TAN이 망막 임상데이터에 대하여 Markov Blanket된 GBN보다 예측정확성이 우수하다. 본 연구에서는 가장 높은 성능을 보인 TAN을 적용하여 망막 질환 요인들의 상호 의존도 관계를 분석하였다.

4.4 TAN으로 질환 요인간 의존도 분석

실험 결과로서 TAN을 적용한 결과를 그림 5와 같이 얻을 수 있다. 비순환유향그래프(DAG)와 조건

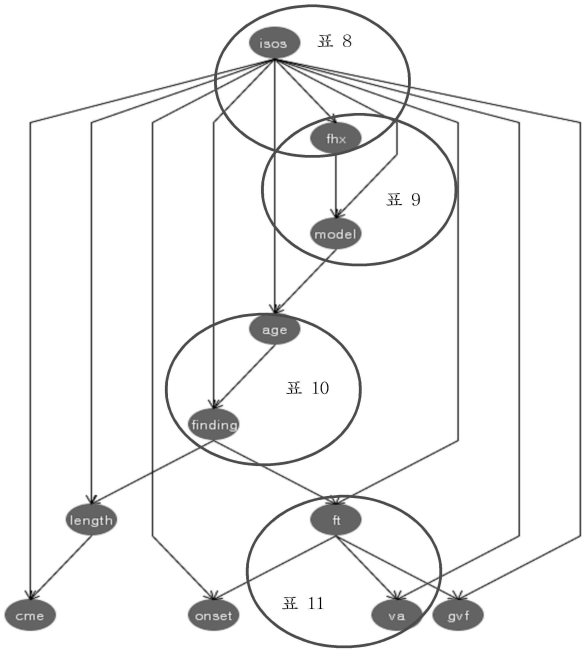

그림 5. 망막 임상데이터에 적용된 TAN

표 5. BN의 분류 성능 비교(단위:%,GBN*:Markov Blanket GBN)

	Training			Test			10-fold
	30	60	90	T30	T60	T90	
TAN	90.0	90.5	91.1	76.7	85.0	90.0	80.0
GBN	78.9	72.2	70.4	65.0	71.7	66.7	74.4
GBN*	76.7	77.8	76.3	58.3	73.3	71.6	75.3

표 6. TAN의 예측정확성 분석(n=90 훈련한 후에 테스트)

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
0	0.85	0.05	0.90	0.85	0.87
1	0.95	0.03	0.95	0.95	0.95
2	0.9	0.08	0.86	0.9	0.88
Average	0.9	0.05	0.90	0.9	0.9
Mean Absolute Error(MAE)			0.14		
Root Mean Squared Error(RMSE)			0.23		

표 7. Markov Blanket된 GBN의 예측정확성 분석(n=90 훈련한 후에 테스트)

Class	TP Rate	FP Rate	Precision	Recall	F-Measure
0	0.7	0.1	0.78	0.7	0.74
1	0.8	0.23	0.64	0.8	0.71
2	0.65	0.1	0.77	0.65	0.7
Average	0.72	0.14	0.73	0.72	0.72
Mean Absolute Error(MAE)			0.28		
Root Mean Squared Error(RMSE)			0.37		

표 8. TAN에서 클래스노드별 fhx(가족력) CPT

isos	가족력없음(0)	가족력있음(1)
0	0.61	0.39
1	0.5	0.5
2	0.61	0.39

부확률표(CPT)를 통하여 여러 요인들 간의 상호의존도를 파악하는 경우에는 TAN이 가장 적합함을 알 수 있다. 예를 들면, 기존연구에서 시력과 망막두께의 상관계수(r)가 [r]=0.576으로 강한 상관관계를 보인다고 연구하였는데[18,19], 이런 결과를 그림 5의 TAN에서 확인할 수 있었다. 그림 5에 나타난 망막질환 요인들의 관계에 대하여 본 연구에서 선택한 TAN으로 표 8, 표 9, 표 10, 표 11에서 분석해 본다.

표 8에서와 같이 클래스노드별 가족력이 미치는 영향을 분석해보면, 가족력이 없는 경우에 망막 질환이 발생할 확률이 0.5~0.61이고 가족력이 있는 경우는 0.39~0.5에 해당한다. 이는 가족력이 없는 상태가 가족력이 있는 경우보다 질환 발생가능성이 높기 때문에 평소에 눈과 망막관리에 신경써야 한다는 것을 알 수 있다. 표 9과 같이 여러 유전 모델의 영향을 분석해 보면, isos의 모든 클래스에서 가족력이 있는 경우(fhx=1)는 유전 모델이 AD(Autosomal Dominant, 우성유전)와 AR(Autosomal Recessive, 열성유전)인 경우가 X(X-linked, 성관련유전)와 S 모델보다 질환 발생 가능성이 높다는 것을 알 수 있다. 가족력이 없는 경우(fhx=0)는 S (Sporadic, 독립적) 모델의 확률 값이 약 97%로 높았으며 대부분 이에 해당하는 것을 알 수 있다.

표 10에서는 TAN을 이용한 결과로 isos 클래스 노드별 나이에 따른 OCT상 소견과의 상호의존관계를 분석했다. 정상군(isos=0) 일 때는 황반에 부종이 생기는 CME(황반부종)이나 망막위에 팽팽한 막이 웃자라는 ERM(망막전막) 등의 증상이 거의 없다. 그러나 부분군(isos=1), 손상군(isos=2)으로 질병이 진행되면서 이들의 증상이 많이 나타나고 있다. 특히, 나이가 많을수록 CME, ERM의 확률 값이 높아지고 있다. 부분군(isos=1)에서 CME가 나타날 확률이 LOW(30.5세 미만)나 MID(30.5이상, 42세 미만)인 경우는 약 14%~16%이며, HI(42세 이상, 56세미만)일 때가 약 61%, HHI(56세이상) 일 때는 약 59%로 나이가 많을수록 확률 값이 커진다. 손상군(isos=2)에서는 나이가 많을수록 CME 뿐 만 아니라, ERM 확률 값도 증가하고 있다. 예를 들면 손상군(isos=2)에서 ERM이 나타날 확률이 나이가 LOW일 때 5.2%, MID 일 때 3.8%, HI 일 때는 9.7% 이고 HHI 일 때는 15.8% 이다. 또한 손상군(isos=2)에서 나이가 HHI인 경우는 CME+ERM 이 같이 나타날 확률이 13.2%

표 9. TAN에서 클래스노드별 fhx(가족력)이 미치는 model(유전모델) CPT

isos	fhx	S	AR	AD	X
0	0	0.974	0.009	0.009	0.009
0	1	0.014	0.392	0.473	0.122
1	0	0.968	0.011	0.011	0.011
1	1	0.074	0.394	0.33	0.202
2	0	0.974	0.009	0.009	0.009
2	1	0.014	0.338	0.554	0.095

로 높다. 이것은 HHI 이상인 경우에 이런 증상들이 같이 발생할 수 있음을 알 수 있다.

표 11의 결과를 분석해 보면, 정상군(isos=0)일 경우에는 망막두께(ft)의 범위에 관계없이 CPT의 최고값이 EYE-1, EYE-2 영역에 나타났다. EYE-1, EYE-2, EYE-3, EYE-4로 갈수록 시력이 나쁜 상태이다. 하지만 부분군(isos=1)과 손상군(isos=2)이 되면서 망막두께는 얇아지게 되고 부종일 때는 두꺼워지면서 시력이 나빠진다는 것을 알 수 있다. 예를 들어 손상군(isos=2)인 경우, 모든 ft 레벨에서 CPT의 최고값이 EYE-3, EYE-4 영역에 나타났다. ONE 레벨일 때는 EYE-3의 분포가 약 60% 이고, FOUR 레벨일 경우에는 EYE-4의 분포가 약 78%로 나타났다. 이런 결과는 기존 연구에서 단순히 시력과 망막두께의 상관계수(r)가 [r]=0.576으로 강한 상관관계를 보

표 10. TAN에서 클래스노드별 age(나이)가 미치는 finding(소견) CPT

isos	age	finding					
		CME	ERM	CME+ERM	MH	taut	NO
0	LOW	0.094	0.032	0.016	0.016	0.016	0.828
0	MID	0.06	0.06	0.015	0.015	0.106	0.742
0	HI	0.218	0.131	0.022	0.022	0.065	0.543
0	HHI	0.056	0.167	0.028	0.028	0.083	0.639
1	LOW	0.143	0.143	0.018	0.018	0.125	0.554
1	MID	0.16	0.04	0.02	0.02	0.38	0.38
1	HI	0.613	0.032	0.016	0.016	0.113	0.21
1	HHI	0.59	0.046	0.023	0.023	0.023	0.295
2	LOW	0.052	0.052	0.026	0.026	0.023	0.605
2	MID	0.038	0.038	0.13	0.019	0.167	0.611
2	HI	0.171	0.097	0.012	0.061	0.305	0.354
2	HHI	0.211	0.158	0.132	0.026	0.079	0.395

표 11. TAN을 이용한 클래스 노드별 ft(망막두께)가 미치는 시력 CPT

isos	ft	EYE-1	EYE-2	EYE-3	EYE-4
0	ONE	0.1	0.5	0.3	0.1
0	TWO	0.192	0.346	0.346	0.115
0	THREE	0.547	0.359	0.047	0.047
0	FOUR	0.712	0.227	0.045	0.015
0	FIVE	0.559	0.324	0.088	0.029
1	ONE	0.438	0.438	0.062	0.062
1	TWO	0.357	0.271	0.357	0.014
1	THREE	0.019	0.712	0.25	0.019
1	FOUR	0.1	0.1	0.7	0.1
1	FIVE	0.135	0.25	0.596	0.019
2	ONE	0.102	0.17	0.602	0.125
2	TWO	0.017	0.293	0.293	0.397
2	THREE	0.071	0.071	0.5	0.357
2	FOUR	0.071	0.071	0.071	0.786
2	FIVE	0.038	0.038	0.577	0.346

임을 표시하였는데[18,19], 이러한 관계를 본 연구에서는 자세하게 표 11에서 보이고 있다.

5. 결론 및 향후 연구

지금까지는 망막 질환 데이터마이닝에서 주로 SPSS와 같은 통계툴을 사용하여 요소들 간의 분산분석, 상관계수와 클래스별 분포 상태 등을 알아볼 수 있었다. 그러나 본 논문에서는 베이지안 네트워크를 망막 임상데이터에 처음으로 적용하여, 각 질환 요인들 간의 상호의존 관계와 의존도의 강도를 분석하였다.

Weka에서 TAN, GBN과 Markov Blanket된 GBN에 대하여 분류성능을 실험한 결과, TAN은 분류성능이 우수하다고 알려진 Markov Blanket된 GBN보다 예측정확률이 높았다. 이는 Markov Blanket으로 특징축소된 GBN이 최소한의 설명 노드들로 축소구성되어, TAN보다 일반화 능력이 떨어지는 결과를 보였다고 판단된다. 높은 성능을 보인 TAN의 비순환유향그래프(DAG)와 조건부확률표(CPT)에 근거한 분석을 통해, 질환 요인들 간의 확률적 의존관계를 파악하여 상호의존도를 알 수 있었다. 예를 들면, 가족력이 없는 상태가 가족력이 있는 경우보다 질환 발생가능성이 높기 때문에, 평소에 눈과 망막관리에 신경써야 한다는 것을 알 수 있었다. 또한 망막이 손상되어 부분군(1)과 손상군(2)이 되면 망막두께가 얇아지거나 부종으로 두꺼워지면서 시력이 나빠진다는 것을 알 수 있었다.

본 연구의 실험결과, 순수 베이지안의 단순함과 네트워크 속성들 사이의 의존성을 표현하는 능력을 결합한 TAN이 높은 예측정확률을 보이며, 질환 요인들 간의 상호의존 관계를 가장 잘 나타내었다. 따라서 망막 질환의 진단과 예후 예측에 정보를 제공하기 위해 TAN의 활용을 우선적으로 제안한다. 앞으로 연구는 객관성을 입증하기 위해 더 많은 속성들을 가진 임상데이터의 수집이 필요하다. 그리고 망막 임상데이터의 분석에 효과적으로 쓰일 수 있는 의료정보시스템을 구축할 예정이다.

참 고 문 헌

[1] 손호선, 이헌규, 조경환, 류근호, 노기용, “심장질환 진단을 위한 베이지안 분류기법,” 한국정보처리학회 춘계학술발표대회 논문집, 제13권, 제1호, pp. 39-42, 2006.

[2] 정용규, 김인철, “베이지안 망에 기초한 불임환자 임상데이터의 분석,” 정보처리학회논문지B, 제9-B권, 제5호, pp. 625-633, 2002.

[3] 이제영, 최영진, “베이지안 네트워크를 활용한 정신장애 질병 섬망의 주요 위험인자와 오즈비,” 한국데이터정보과학회지, 제22권, 제2호, pp. 217-225, 2011.

[4] N. Friedman, M. Geiger, and M. Goldszmidt, “Bayesian Network Classifiers,” *Machine Learning,* Vol. 29, Issue 2-3, pp. 131-163, 1997.

[5] Liangxiao Jiang, Zhihua Cai, Dianhong Wang, and Harry Zhang, “Improving Tree Augmented Naive Bayes for Class Probability Estimation,” *Konwledge-Based Systems*, Vol. 26, pp. 239-245, 2012.

[6] A. Chinnasamy, W.K. Sung, and A. Mittal, “Protein Structure and Fold Prediction using Tree-Augmented Naive Bayesian Classifier,” *Pacific Symposium on BioComputing* 9, pp. 387-398, 2004.

[7] J. Cheng and R. Greiner, “Comparing Bayesian Network Classifiers,” *Proc. of the 15th Conf. on Uncertainty in Artificial Intelligence*, pp. 101-107, 1999.

[8] I. Tsamardinos, C.F. Aliferis., and A. Statnikov, “Time and Sample Efficient Discovery of Markov Blankets and Direct Causal Relations,” *The 9th ACM SIGKDD International Conference on Konwledge Discovery and Data Mining*, pp. 24-27, 2003.

[9] 유형곤, 유전성 망막질환, 서울대학교출판문화원, 서울, 2011.

[10] S. Aizawa, Y. Mitamura, T. Baba, A. Hagiwara, K. Ogata, and S. Yamamoto, “Correlation Between Visual Function and Photoreceptor Inner/outer Segment Junction in Patients with Retinitis Pigmentosa,” *Eye*, Vol. 23, No. 2, pp. 304-308, 2009.

[11] A. Oishi, A. Otani, M. Sasahara, H. Kojima,

H. Nakaura, M. Kurioto, and N. Yoshimura, "Photoreceptor Integrity and Visual Acuity in Cystoids Macular Oedema Associated with Retinitis Pigmentosa," *Eye*, Vol. 23, No. 6, pp. 1411-1416, 2009.

[12] Michael A. Sandberg, Robert J. Brockhurst, Alexander R. Gaudio, and Eliot L. Berson, "The Association Between Visual Acuity and Central Retinal Thickness Retinitis Piagmentosa," *IOVS*, Vol. 46, No. 9, pp. 3349-3354, 2005.

[13] 김현미, 우용태, 정성환, "망막색소변성 데이터의 예후예측을 위한 패턴 분류," 멀티미디어학회논문지, 제15권, 제6호, pp. 701-710, 2012.

[14] Jensen F.V., *An Introduction to Bayesian Networks*, Springer-Verlag, New York, 1996.

[15] P. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Addison Wesley Longman, California, 2007.

[16] J. Cheng and R. Greiner, "Learning Bayesian Belief Network Classifiers : Algorithm and System," *Proc. the Fourteenth Canadian Conference on Artificial Intelligence,* No. 2056, pp. 141-151, 2001.

[17] Weka Manual 3.6.5, http://www.cs.waikato.ac.nz/ml/weka, 2012.

[18] Michael A. Sandberg, Robert J. Brockhurst, Alexander R. Gaudio, and Eliot L. Berson, "Visual Acuity Is Related to Parafoveal Retinal Thickness in Patients with Retinitis Pigmentosa and Macular Cysts," *IOVS*, Vol. 49, No. 10, pp. 4568-4572, 2008.

[19] Hyewon Chung, Jong-Uk Hwang, June-Gone Kim, and Young Hee Yoon, "Optical Coherence Tomography in the Diagnosis and Monitoring of Cystoid Macular Edema in Patients with Retinitis Pigmentosa," *Retina*, Vol. 26, No. 8, pp. 922-927, 2006.

김 현 미

1994년 덕성여자대학교 전자계산학과 이학사
2001년 서강대학교 정보시스템전공 공학석사
2012년 창원대학교 컴퓨터공학과 공학박사

1994~2000년 성우그룹 성우전자, 델파이코리아 전산실
2002~2004년 창원문성대학 컴퓨터정보처리과 연구교수
2005년~현재 창원대학교, 마산대학 등 강사
관심분야 : 기계학습 및 패턴인식, 데이터마이닝

정 성 환

1979년 경북대학교 전자공학과 (공학사)
1983년 경북대학교 대학원(공학석사-정보통신)
1988년 경북대학교 대학원(공학박사-영상처리)

1983년~1985년 한국전자통신연구소 연구원
1986년 전자계산기기술사, 1992년 정보처리기술사
2003년 정보시스템감리사
1993~1994년 Univ. of California(UCSB) Post-Doc.
1999~2000년 Colorado School of Mines 연구교수
2008~2009년 Univ. of Missouri(UMKC) 방문교수
1988년~현재 창원대학교 컴퓨터공학과 교수
관심분야 : 영상처리, 컴퓨터비전 및 패턴인식, 멀티미디어 정보보호