

망막색소변성 데이터의 예후 예측을 위한 패턴 분류

김현미[†], 우용태^{**}, 정성환^{***}

요 약

망막색소변성(RP: Retinitis Pigmentosa)이란 가장 흔한 유전성 망막질환이다. 정상적인 사회활동을 영위 하던 사람들이 이 질병으로 시력이 손상되면서 좌절과 고통을 겪는다. 또한 국가적 차원에서 이들의 경제 활동이 끊김에 따라 경제활동 인구 감소에 따른 손실 또한 크다고 하겠다. 이에 망막색소변성 질환에 대한 임상 예후 정보를 제공할 수 있는 연구기반이 절실히 요구되고 있다. 본 연구는 망막색소변성 데이터에 대한 패턴 분류를 통해 예후 예측이 가능함을 제안한다. 기존에는 주로 SPSS등을 활용한 통계 처리 결과가 데이터 분석에 적용되었다. 그러나 본 연구에서는 기계학습과 자동 패턴 분류를 실험하였다. SVM(Support Vector Machine)과 여러 다양한 패턴분류기들을 실험을 위해 사용하였다. 제안한 방법은 SVM 분류기에 의하여 RP 데이터가 자동적으로 분류된 결과를 바탕으로 예후 예측이 가능함을 확인하였다.

Pattern Classification of Retinitis Pigmentosa Data for Prediction of Prognosis

Hyun-Mi Kim[†], Yong Tae Woo^{**}, Sung-Hwan Jung^{***}

ABSTRACT

Retinitis Pigmentosa(RP) is a common hereditary disease. While they have been normally living, those who have this symptom feel frustration and pain by the damage of visual acuity. At the national level, the loss of the economic activity due to the reduction of economically active population will be also greater. There is an urgent need for the base study that can provide the clinical prognosis information of RP disease. In this study, we suggest that it is possible to predict prognosis through the pattern classification of RP data. Statistical processing results through statistical software like SPSS(Statistical Package for the Social Service) were mainly applied for the conventional study in data analysis. However, machine learning and automatic pattern classification was applied to this study. SVM(Support Vector Machine) and other various pattern classifiers were used for it. The proposed method confirmed the possibility of prognostic prediction based on the result of automatically classified RP data by SVM classifier.

Key words: Prediction of Prognosis(예후 예측), Pattern Classification(패턴 분류), SVM(Support Vector Machine)(서포트 벡터 머신), RP(Retinitis Piagmentosa)(망막색소변성), SPSS(Statistical Package for the Social Service)

※ 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 : 2011년 9월 15일, 수정일 : 2012년 1월 25일
완료일 : 2012년 3월 18일

[†] 준회원, 창원대학교 컴퓨터공학과
(E-mail: hmkim@changwon.ac.kr)

^{**} 준회원, 창원대학교 컴퓨터공학과
(E-mail: ytwoo@changwon.ac.kr)

^{***} 종신회원, 창원대학교 컴퓨터공학과

1. 서 론

망막색소변성이란 가장 흔한 유전성 망막질환으로, 안구내 신경조직인 망막의 광수용체층이 유전적 결함에 의해 변성이 발생하게 된다. 증상은 초기에는 야맹증으로 시작하여 점차 시야가 좁아져 결국에는 실명까지 이르게 된다. 정상적인 사회활동을 영위하던 사람들에게 시력과 시야가 손상되면서 느끼는 좌절과 고통이 클 뿐 아니라 국가적 차원에서도 이들의 경제 활동이 끊김에 따른 경제활동 인구 감소에 따른 손실 또한 크다고 하겠다. 따라서 망막색소변성 질환에 대한 전반적인 예후(prognosis)¹⁾를 알려주는 임상 예후에 대한 정보를 제공할 수 있는 연구기반이 중요하게 요구되고 있다[1].

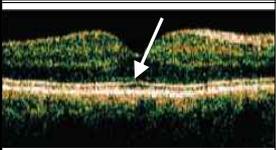
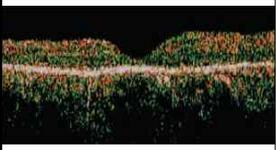
이전의 연구들에서는 표 1과 같이 망막광수용체층이 변성으로 파괴되면 광수용체의 두 층인 내절·외절의 경계선(IS/OS: junction of inner segment and outer segment of photoreceptor layer)이 소실되거나 끊어져 이미지에서 안보이게 되고, 이런 환자들은 시력이 불량하다는 것이 알려졌다. 이전 연구들에서는 SPSS(Statistical Package for the Social Science)와 같은 통계툴을 이용하여 통계적으로 의미 있는 결과를 도출하였다[2-6].

그러나 SPSS 통계툴을 활용하기 전에 전문가가

직접 진단하고 데이터를 기준(IS/OS의 존재 여부)에 따라 표 1의 세 클래스로 분류하는 절차가 반드시 필요하였다. 이에 본 연구에서는 전문가가 직접 진단하여 분류 클래스를 결정하지 않아도 기계학습으로 자동 분류가 가능한지를 실험한다. 이에 최근에 우수한 분류성능으로 얼굴 인식, 문자 인식, 의학, 생물학, 유전학 등 여러 분야에서 각광받고 있는 SVM(Support Vector Machine)을 비롯하여 MLP(Multi-Layer Perceptron), Logistic 회귀분석, 의사결정트리(decision tree) 등 여러 알고리즘들을 적용하여 분류 실험을 한다. 그리하여 높은 분류 성능을 보인 패턴분류기로 망막색소변성 데이터를 학습시키고 자동 분류를 가능하게 하여 그 결과를 산출함으로써 SPSS 분석을 이전보다 신속하게 할 수 있다. 그 뿐만 아니라 망막색소변성 질환의 예후 예측에 도움을 주고자 자동 패턴 분류기의 활용을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존 연구를 살펴보고 본 연구에서 제안하는 시스템을 기술한 다음에 관련된 지식을 알아본다. 관련된 지식으로 현재 패턴 인식 분야에서 널리 쓰이고 있는 SVM에 대해 알아보겠다. 3장에서는 망막색소변성 데이터 분류 실험에 관한 실험 방법과 결과를 설명한다. 그리고 예후 예측에 대한 토의를 거쳐, 4장에서 결론을 맺고 향후 연구를 기술한다.

표 1. IS/OS 상태와 시력과의 관계(IS/OS: 망막광수용체층의 경계선, 상단 화살표)

OCT 사진	IS/OS 상태	시력
	IS/OS 경계선이 끊어지지 않고 연속적으로 존재	대체로 양호
	IS/OS 경계선이 부분적으로 끊어지고 불연속적으로 존재	불량이거나 나빠지기 시작함
	IS/OS 경계선이 끊어져서 안보임	시력 불량하고 악화되면 실명 가능성 있음

1) 예후[prognosis]는 병세 등의 진행 및 회복에 관한 예측을 위한 사전 판단을 뜻하는 용어(병세와 관련하지 않은 기계의 고장 진단에도 널리 쓰임)

2. 관련 연구

2.1 기존 연구

지금까지 보고된 기존 연구에서는 망막색소변성 임상 데이터를 가지고 주로 SPSS 통계 처리를 통한 분석 방법을 사용하였다. 먼저 전문가의 진단으로 측정 기준인 IS/OS(망막광수용체층의 경계선)의 상태에 따라 IS/OS 라인(line)이 끊어짐 없이 완전히 관찰되는 정상그룹, 불연속적으로 존재하는 부분그룹, 라인이 손상되어 완전히 소멸되어 있는 손상그룹의 세 그룹으로 분류한다. 연구결과를 살펴보면 Michael[2], Aizawa[3]와 Oishi[4] 등은 IS/OS 등급은 시력(VA: Visual Acuity)과 중심망막두께(CFT: Central Foveal Thickness)와 관련하여 통계적으로 유의한 의미관계가 있다고 주장하였다.

수집한 임상 데이터를 가지고 실제로 기존 연구와 같은 실험방법을 시도하였다. SPSS 통계툴을 이용

표 2. IS/OS 클래스와 각 요소간의 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 0.05수준에서 유의, **상관계수는 0.01수준에서 유의

한 상관분석 처리결과는 표 2와 같다. 상관계수(r)의 값이 $0.3 < |r| < 0.7$ 이면 뚜렷한 상관관계를 의미한다. IS/OS와 중심망막두께는 -0.402 이고 IS/OS 클래스와 logMAR²⁾ 시력은 $+0.645$ 의 측정치가 나왔다. 여기서 IS/OS 클래스는 중심망막두께보다는 logMAR시력에서 더 유의함을 알 수 있다.

Aizawa[3]는 각 특징 요소 중에서 IS/OS 클래스를 독립변수로 나이, 시력, 중심망막두께를 종속변수로 하여 회귀분석 결과를 알아보았다. 여기에서는 IS/OS 클래스와 나이와는 약하게 유의한 관계가 있고 IS/OS 클래스와 망막중심두께, IS/OS 클래스와 시력과는 매우 유의한 상관관계가 있음을 보이고 있다. Osishi[4]도 IS/OS 등급이 RP환자에게서 시력을 위한 가장 중요한 요소임을 설명하고 있다.

Hyewon Chung[5]과 Michael[6] 등은 시력과 중심망막두께는 통계적으로 의미 있는 결과를 가지며 이를 뒷받침할 수 있는 회귀방정식을 구하였다. 수집한 임상데이터로 실제로 적용한 회귀방정식은 그림 1과 같다. 이로써 시력은 망막중심두께와의 상관관계가 있다는 것을 알 수 있다.

실제 SPSS 통계툴을 적용한 실험 결과로 IS/OS 라인의 변성이 심할수록 중심망막두께도 감소하거나 부종으로 두꺼워지게 되고 결국 시력도 떨어지게 됨을 알 수 있었다[2-4]. 그러므로 RP를 가진 환자를 모니터링 할 때 IS/OS 상태가 어느 요소보다도 더욱

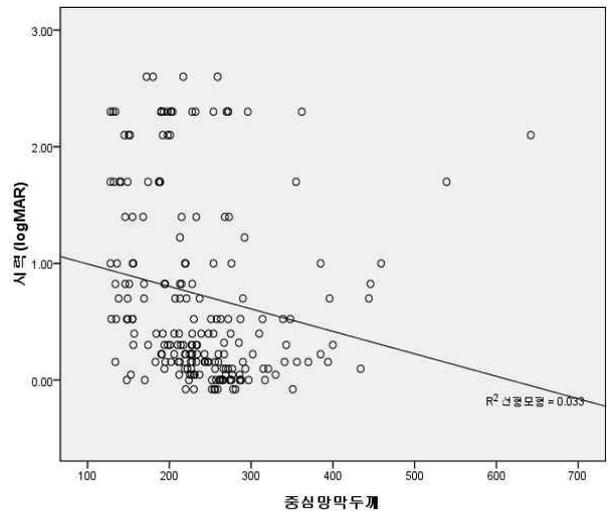


그림 1. 중심망막두께와 시력의 회귀방정식($Y = -0.02X + 1.187$)

중요한 설계 요소가 되어야 한다. 또한 망막색소변성이 아닌 다른 의료데이터 분야에서 패턴 분류가 연구된 예는 다음과 같다. 이광기[7]는 UCI 데이터베이스에 있는 간, 당뇨, 유방암 데이터를 가지고 혼용 유전자 알고리즘 기반의 분류 방법을 동원함으로써 의료진단 문제와 같은 다차원의 패턴 분류 문제를 해결할 수 있는 방안을 제시하였다. 이와 같이 의료데이터를 위한 자동 패턴 분류 문제는 망막색소변성 질환에서도 적용되어야 한다. 이를 위하여 SVM과 같은 기계학습을 통한 훈련으로 자동 분류를 가능하게 함으로써 IS/OS 클래스별 분포 상태가 자동으로 파악되어 치료 후 효과를 신속하게 예측할 수 있게 된다.

2) logMAR 시력은 대수시력표에서의 시력을 말함. logMAR 시력 = $-\log(\text{시력검사표 시력})$

2.2 제안시스템

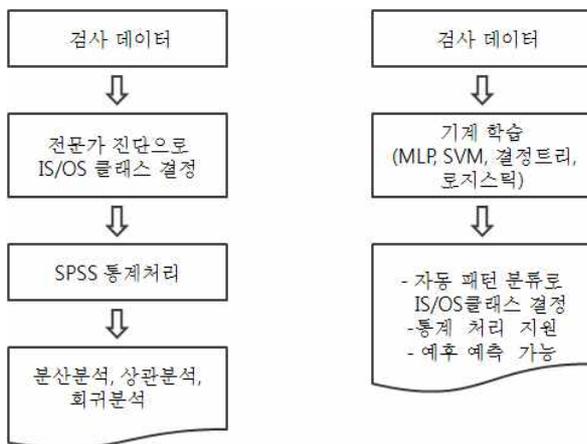
2.2.1 시스템 흐름도

기존 시스템과 제안하는 시스템을 그림 2와 같이 도식화하였다. 기존의 프로세스를 간소화하고 자동 분류를 가능하게 하였을 뿐만 아니라 치료 후 시력 예후를 가늠해보는 추론 기능까지 적용될 수 있다.

기존 시스템에서 전문가들은 빛간섭 단층촬영(OCT: Optical Coherence Tomography) 검사상의 데이터를 가지고 직접 진단하여 IS/OS 클래스를 결정한다. 이러한 측정값들을 텍스트 파일로 변환한다. 그런 다음, SPSS 통계툴에 적용하여 요소들 간의 상관관계, 회귀분석과 클래스별 분포 상태 등을 알아볼 수 있었다[2-6].

본 논문에서 제안하는 시스템은 그림 2의 (b)와 같이, 일단 훈련 데이터로 기계학습이 이루어진 상태에서는 매번 전문가가 직접 진단하여 IS/OS의 클래스를 결정할 필요가 없다. 새로 검사한 OCT 데이터(시력검사 포함)가 있을 때마다 텍스트 파일로 생성하고, 이를 패턴 분류기에 입력하여 자동 분류시킨다. 그리고 그 결과물로 통계 처리를 지원하고 예후 예측에 사용한다.

제안 시스템의 잇점을 예로 들면, 부분군 클래스에 속하던 데이터들이 손상군 클래스로 얼마정도 분포되었는가를 알 수 있고 변화의 추이를 신속하게 파악할 수 있다. 이런 연구의 목적을 위한 패턴 분류 알고리즘으로 SVM을 제안한다. 그 이유는 최근에 우수한 분류 성능으로 얼굴 인식, 문자 인식, 의학,



(a) 기존시스템 (b) 제안시스템

그림 2. 기존시스템과 제안시스템의 흐름도

생물학, 유전학 등 여러 분야에서 각광받고 있기 때문이다[8,9]. 이러한 SVM의 우수성을 입증하기 위하여 MLP, Logistic 회귀분석, 의사결정트리 등 여러 알고리즘들과의 분류 실험을 하여 그 성능을 비교해 본다.

2.2.2 SVM

기계학습과 데이터마이닝을 위한 표준 도구들의 하나인 SVM은 인공 지능의 한 분야로 신경망, 베이저안 네트워크(Bayesian network), k-최근접 이웃(k-nearest neighbor), k-평균(k-means) 등과 같은 학습 알고리즘이다. 1979년 V. N. Vapnik에 의해 개발된 통계적 학습 이론(SLT: Statistical Learning Theory)으로 일반적인 통계적 학습 방법에서의 경험적 리스크 최소화(ERM: Empirical Risk Minimization)와는 다른 구조적 리스크 최소화(SRM: Structural Risk Minimization) 원칙에 기반을 두어 더욱 일반화 된 성능을 갖는다[10]. SRM의 개념은 클래스 분리의 마진을 최대화시키는 것이며 이것은 학습 이론을 단지 이론적 분석뿐만 아니라 실세계 문제에서도 적용되는 알고리즘이 될 수 있도록 한다[11].

SVM에서 각각의 object(○, ×)는 N차원 공간에서 한 점에 대응하는 N정수의 벡터 x_i 로 기술된다. 그림 3은 간단한 하이퍼평면(hyperplane)에 의해 두 집단을 분류하는 것에 대한 설명이다. 활성 그룹에 있는 object들은 각각 $y_i = +1$ 로 할당이 되고, 비활성 그룹에 있는 object들은 $y_i = -1$ 로 할당된다. 그림 3에서 이 object들은 다음과 같이 분류된다.

$$w \cdot x_i + b \geq +1 \text{ for } y_i = +1 \text{ (class 1)} \tag{1}$$

$$w \cdot x_i + b \leq -1 \text{ for } y_i = -1 \text{ (class 2)} \tag{2}$$

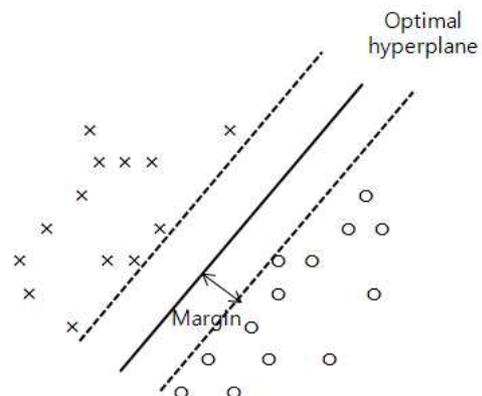


그림 3. SVM의 개념

여기서 w 는 하이퍼평면에 대한 법선 벡터(normal vector)이고, b 는 스칼라 값이다. SVM은 수식 (3)과 같은 조건부 최적화문제의 해를 구함으로써 최대마진을 가지는 최적의 분리가능한 하이퍼평면을 찾는다. 이러한 하이퍼평면에서 가장 가까이에 있는 오브젝트를 지지 벡터(support vector)라고 한다[12].

$$\text{Max}_{w,b} \frac{2}{\|w\|} \text{ subjects to } y_i(w \cdot x_i + b) - 1 \geq 0 \quad (3)$$

여기서 $\frac{2}{\|w\|}$ 는 마진(margin)이다. 위 식은 라그랑제 승수(Lagrange multiplier)를 도입하여 해를 구할 수 있다. 라그랑제 함수로 표현한 조건부 최적화 문제는 보통 KKT(Karush Kuhh Tucker) 조건을 이용하여 푼다. 그리하여 분류결정함수는 다음과 같이 구해진다.

$$\begin{aligned} f(x) &= \text{sgn}(w \cdot x_i + b) \\ &= \text{sgn}\left[\sum_{i=1}^n \alpha_i y_i (x \cdot x_i) + b\right] \end{aligned} \quad (4)$$

이 식은 하이퍼평면이 점들을 두 개의 그룹으로 정확히 분리할 수 없는 비선형 분리 경우로 확장된다. 이런 경우 벡터 내적을 커널 함수 계산으로 대체하여 계산한다. 이때 표 3과 같은 다양한 종류의 커널(kernel) 함수들을 적용할 수 있다[13].

본 연구의 실험에는 일반적인 방법에서 우선적으로 선택하는 커널인 RBF(radial basis function) 커널을 사용한다. 그 이유는 시그모이드(sigmoid) 커널은 벡터 내적의 계산이 파라미터 값에 따라 불분명(0, 또는 ∞)하게 되어 정확성이 RBF 커널보다 떨어지게 된다. 그리고 폴리노미얼(polynomial) 커널은 다른 커널들보다 더 많은 hyperparameter를 가지게 되어 계산적 어려움과 복잡성이 커지게 된다. 그에 비해 RBF 커널은 위의 두 커널들의 단점을 극복할 수 있으며 클래스 라벨(label)과 특징 요소들 사이의 관계가 비선형인 경우라도 적용될 수 있기 때문이다[14].

표 3. 커널 함수의 종류(γ, r, d : 커널 파라미터들)

함 수 식	커널 함수
$k(x_i, x_j) = x_i \cdot x_j$	linear
$k(x_i, x_j) = (\gamma x_i \cdot x_j + r)^d, \gamma > 0$	polynomial
$k(x_i, x_j) = \tanh(\gamma x_i \cdot x_j + r)$	sigmoid
$k(x_i, x_j) = \exp(-\gamma \ x_i - x_j\ ^2), \gamma > 0$	RBF

3. 실험 및 결과

3.1 실험 환경

실험을 위한 시스템 사양은 윈도우즈 7, CPU는 인텔 코어 i5 2.53GHz, RAM은 3GB 사양에서 Matlab 7.1로 실험하였다. 패턴 분류기로는 최근에 개선된 SMO(Sequential Minimal Optimization) 알고리즘으로 구현되어 널리 쓰이고 있는 LIMSVM을 적용하여 분류 실험을 하였고 MLP, Logistic 회귀분석, 의사결정트리 등과의 분류 성능 비교를 위하여 WEKA에서 실험하였다.

표 4. 실험 사양

Item	Specification
CPU	Intel Core i5 2.53GHz
RAM	3GB
OS	Windows 7
Language	MATLAB 7.1
Data	OCT → text
Classifiers	LIMSVM, WEKA

3.2 실험 방법

A대학병원 안과학교실의 147명의 망막 임상데이터를 가지고 우안, 좌안 OCT 데이터를 각각 독립적으로 보아 총 294안의 샘플을 확보하였다. IS/OS(망막광수용체층의 경계선)의 상태에 따라 다음과 같이 3개의 클래스로 분류하였다.

- 정상군(ω_0): 망막광수용체층의 경계선이 완전하게 존재함
- 부분군(ω_1): 망막광수용체층의 경계선이 불연속적으로 존재함
- 손상군(ω_2): 망막광수용체층의 경계선이 존재하지 않음

정상군인 경우에는 IS/OS의 라인이 완전하게 존재하는 경우이며 시력이 좋다. 부분군, 손상군으로 갈수록 시력이 저하되고 망막이 손상되면서 망막두께도 얇아지거나 두꺼워진다[5-6]. 총 294안 중에서 정상군 140안, 부분군 40안, 손상군 114안으로서 사전 확률은 각각 $P(\omega_0) = 0.48, P(\omega_1) = 0.13, P(\omega_2) = 0.39$ 이다.

표 5. 특징 요소 추출

	Age	CFT	IS/OS	logMAR
1	52.0	444	1	0.70
2	52.0	385	1	0.22
3	50.0	339	0	0.52
4	67.0	217	0	0.30
5	43.0	321	0	0.10
6	34.0	261	0	0.00
7	31.0	362	2	2.30
8	66.0	446	1	0.82
9	52.0	343	0	0.30
10	48.0	642	2	2.10
11	20.0	278	0	-0.08
12	36.0	155	2	1.00

특징 요소 추출은 기존 논문과 전문가의 조언을 참조하여 표 5와 같이 나이(Age), 중심망막두께(CFT), IS/OS클래스, logMAR 시력을 들 수 있다.

수집한 샘플을 가지고 SVM 등의 다양한 패턴 분류기를 적용하는 실험에서 사용할 망막색소변성 데이터를 훈련 집합(training set)과 테스트 집합(test set)으로 구분하였다. 훈련 집합 크기(n)는 각 클래스별로 20부터 100까지 10씩 증가시키면서 실험하였다. 훈련집합이 n=20, 30인 경우에는 부분군의 데이터 크기가 40으로 적어서 모델안정성과 검증 테스트(validation test)를 위하여 k-fold 교차 검증을 도입하였다[13]. 즉 정상군, 부분군, 손상군 모두 표 6과 같은 방법으로 k-fold 교차 검증 기법을 사용하였다.

n=40 이상인 경우에는 정상군, 부분군, 손상군 모두 훈련 집합 크기별로 2개의 훈련 집합을 구성하였고 테스트 집합도 2개로 구성하였다. 그리하여 훈련 집합이 다름에 따라 서로 다른 테스트 집합으로 실험하였다. 이때 부분군은 데이터를 랜덤으로 중복하여 재샘플링(resampling) 기법으로 훈련 집합과 테스트 집합을 구성하였다.

표 6. 세 클래스별 실험 방법(*k-fold 교차 검증 실험)

클래스 \ 훈련집합 크기	*N=20 (k=2)	*N=30 (k=4)	N=40,50,60,70,80,90,100 (훈련집합 2 set)
정상(ω_0)	테스트 2 set	테스트 4 set	테스트 2 set
부분(ω_1)	테스트 2 set	테스트 4 set	테스트 2 set
손상(ω_2)	테스트 2 set	테스트 4 set	테스트 2 set

3.3 실험 결과

본 연구에서 실험한 SVM을 비롯한 MLP, Logistic 회귀분석, 의사결정트리 J48 등 여러 알고리즘들과의 분류 실험 결과는 그림 4에 나타났다. Meyer[15]는 16개의 분류 알고리즘을 SVM과 비교한 결과를 통하여 SVM은 소수의 경우를 제외하고는 다른 알고리즘에 비해 같던지 우수하다는 결론을 제시하였는데 본 실험에서도 좋은 결과를 보이고 있다.

Logistic 회귀분석과 MLP는 훈련집합의 크기와 크게 상관없이 대체로 완전한 분류 성능을 보였다. 그러나 의사결정트리 J48과 SVM은 훈련집합의 크기가 증가할수록 분류 정확률이 높게 나타나는 것을 알 수 있다. 특히, SVM은 훈련집합 크기가 n=80이상일 때 의사결정트리 J48보다 높은 정확률을 보이며 n=100일 때는 97.5%의 신뢰할만한 성능을 보였다. 망막색소변성 데이터와 같은 의료 데이터 영역에서는 SVM이 훈련집합 크기가 일정한 수준 이상일 때 (n=90) 뛰어난 성능을 보임을 확인하였다.

또한 데이터의 크기가 작아서 재샘플링 기법을 적용한 부분군(ω_1) 일 때에도 분류 성능이 높아짐을 알 수 있다. SVM에서는 특히 n=90 이상인 경우에 평균 92.5%의 높은 분류 정확률을 보인다. 이는 실험 환경

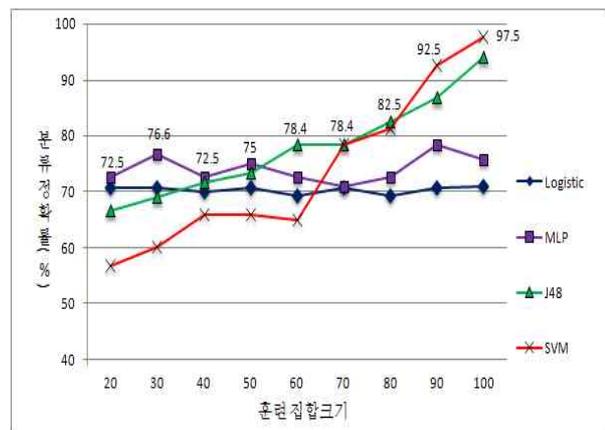


그림 4. 분류 성능 비교

인 LIMSVM에서 다중클래스 분류 알고리즘으로 ‘일-대-일(one-against-one)’을 구현하고 있는데, 다른 멀티클래스 알고리즘인 ‘일-대-다(one-against-all)’ 보다는 실제 적용에 있어서는 ‘일-대-일’과 DAGSVM (Directed Acyclic Graph SVM) 방법이 더욱 적합하다고 주장하고 있다[16,17]. 즉, M개의 클래스가 있다면 ‘일-대-다’는 분류함수가 M개이지만 ‘일-대-일’은 $M(M-1)/2$ 개만큼 존재하여 더욱 세밀하게 구분되어 입력 벡터의 인식률을 높인다고 판단된다.

표 7과 표 8은 SVM을 적용하여 훈련 집합 $n=40$ 과 $n=100$ 의 분류 결과를 클래스별로 분포 상태를 보여 준다. 이를 바탕으로 훈련집합 크기를 $n=100$ 이상으로 하고 주기적인 재검사에 SVM 분류기를 활용하면 질병 진행 정도를 알아낼 수 있다. 그리고 이전과 현재의 변화된 추이를 신속하고 정확하게 알아볼 수 있어서 예후 예측에 도움이 될 수 있다.

3.4 예후 예측을 위한 토의

패턴 분류기를 구현하여 얻을 수 있는 잇점으로 예후 예측의 가능성의 예를 보여주기 위한 실험을 하였다. 이를 위하여 먼저 A대학병원 안과학 교실로부터 기존 검사한 동일 환자에 대하여 재검사한 데이터를 일부 확보하였다. 표 9는 최초 진단시의 데이터와 치료 후 일년 정도 지나서 다시 진단한 시력 검사, OCT 데이터 특징들을 보인다. 정상군인 환자 A는 일년 후에도 클래스간 변화가 없었으나 환자 B의 좌안은 중심망막두꺼(CFT)가 얇아지고 시력이 나빠지면서 부분군에서 손상군 클래스로 질병이 진행되었다는 것을 패턴 분류기 실험으로 알 수 있다. 본 연구에서 실험하여 신뢰할 만한 분류 성능을 보인 SVM과 같은 패턴 분류기를 이용한 예후 예측 방법을 그림 5와 같이 제시하였다. 전문가의 조언으로는 주로 단일군, 혼합군과 같은 방법으로 예후 예측이 가능하다. 예를 들자면 정상군(ω_0), 부분군(ω_1), 손상

표 7. 훈련집합 $n=40$ 일때 분류 분포(단위: %, T_1, T_2 : 서로 다른 테스트 집합)

테스트 집합		정 상(ω_0)		부 분(ω_1)		손 상(ω_2)		평 균	
		T_1	T_2	T_1	T_2	T_1	T_2	부분제외	전 체
SVM	정상(ω_0)	35	40	5	·	·	·	68.8	65.9
	부분(ω_1)	·	15	50	70	·	·		
	손상(ω_2)	65	45	45	30	100	100		
정 확 률		37.5		60		100			

표 8. 훈련집합 $n=100$ 일때 분류 분포(단위: %, T_1, T_2 : 서로 다른 테스트 집합)

테스트 집합		정 상(ω_0)		부 분(ω_1)		손 상(ω_2)		평 균	
		T_1	T_2	T_1	T_2	T_1	T_2	부분제외	전체
SVM	정상(ω_0)	100	95	5	·	·	·	98.7	97.5
	부분(ω_1)	·	5	95	95	·	·		
	손상(ω_2)	·	·	·	5	100	100		
정 확 률		97.5		95		100			

표 9. 재검사 데이터로 SVM 분류를 통한 예후 예측의 예

Patient	년도	1차				2차				IS/OS 변화 추이
		나이	CFT	시력	IS/OS	나이	CFT	시력	IS/OS	
A	우	50	372	0.15	정상	51	383	0.1	정상	정상→정상
	좌	50	316	0.3	정상	51	319	0.3	정상	정상→정상
B	우	36	212	0.4	손상	37	194	0.5	손상	손상→손상
	좌	36	202	0.7	부분	37	197	0.4	손상	부분→손상

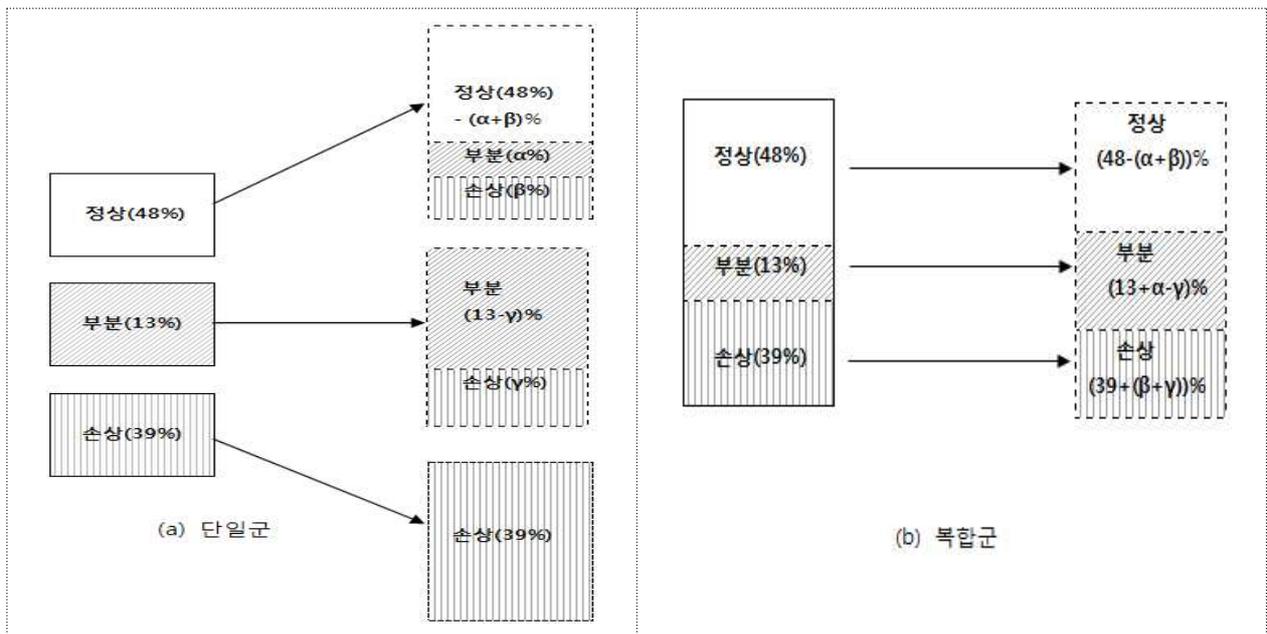


그림 5. 예후 예측을 위한 데이터의 추이

군(ω_2)에 속했던 환자를 진단하고 치료한 다음, 일정 기간 이후에 상태 변화를 알기 위해 다시 검사하게 된다.

이전 정보대로 각 클래스별 단일군으로 새로운 입력 데이터를 구성한다. 그리하여 이미 학습되어진 분류기에 넣으면, 그림 5의 (a)와 같이 현재 상태의 변화 추이를 알 수 있게 된다. 전문가에 의한 진단과 결과 판정을 기다리지 않고 SPSS 통계처리를 적용하지 않아도 신속하게 질병이 어느 정도 진행되었음을 추이할 수 있다.

즉 정상군에서 부분군으로 α 만큼, 손상군으로 β 만큼 진행되어 전체 $-(\alpha+\beta)\%$ 의 변화가 있었고 부분군에서는 $+(\alpha-\gamma)\%$ 의 변화를 보이는데 이는 손상군 클래스로 정상군과 부분군에서 $+(\beta+\gamma)\%$ 만큼 질병이 진행되었음을 알 수 있다. 또한 그림 5의 (b)와 같이 전체 분포 상태의 추이를 파악할 수 있다. 이런 경우는 모든 클래스의 환자를 대상으로 치료 후 상태가 호전되거나 악화되었는가를 알아보기 위한 방법으로 정상군(ω_0), 부분군(ω_1), 손상군(ω_2)의 데이터를 모두 포함한 혼합군으로 입력 데이터를 구성하여 분류 실험을 한다. 혼합군은 전체 통계 자료를 나타낼 때 유용하게 쓰일 수가 있다.

4. 결 론

본 연구에서는 망막색소변성 데이터를 가지고 최근에 우수한 성능으로 각광받고 있는 SVM을 비롯하여 MLP, Logistic 회귀분석, 의사결정트리 J48 등

여러 알고리즘을 적용하여 데이터 분류 실험을 하였다. 훈련집합의 크기가 클수록 SVM과 의사결정트리의 분류 정확률이 높아졌고 데이터 크기가 $n=90$ 이상일 때 SVM이 의사결정트리보다 더욱 신뢰할 만한 성능($n=100, 97.5\%$)을 보였다.

이전 연구의 SPSS와 같은 통계 처리 방법으로는 전문가가 새로 검사한 데이터를 일일이 진단하여 IS/OS(망막광수용체층의 경계선) 클래스를 결정하여야만 단일군, 혼합군의 분포 상태를 알 수 있었다. 그러나 본 연구에서는 기계학습을 통한 패턴 분류기를 활용함으로써 각 클래스별 단일군으로 또는 혼합군으로 신속하게 새로운 분포 상태를 알 수 있었다. 그리하여 자동으로 최상위 클래스인 IS/OS 클래스를 결정함으로써 기존시스템의 통계 처리를 신속하게 돕고, 질병의 진행 속도를 빠르게 알아볼 수 있었다.

본 연구에서 제안하는 자동 패턴 분류기를 활용하여 의료데이터와 같은 다차원의 패턴 분류 문제를 해결할 수 있는 방안을 제시하였다. 기존 시스템과 병행하여 제안 시스템을 활용하면 좀 더 신속하고 효율적으로 진단 및 치료 후 효과를 예측하는 데에 사용될 수 있겠다.

앞으로의 과제는 여러 주요 속성들 간의 상호의존성을 표현하고 이것을 바탕으로 분류 클래스를 비교적 정확하게 예측할 수 있는 베이저안 네트워크(Bayesian network)같은 데이터마이닝[18]에 대하

여 연구할 계획이다. 또한 자동화된 패턴 분류기로 망막색소변성의 예측 예측에 적용을 위한 풍부한 임상 데이터의 확보가 요구된다.

참 고 문 헌

- [1] 유형곤, 유진성 망막질환, 서울대학교출판문화원, 서울, 2011.
- [2] Michael A. Sandberg, Robert J. Brockhurst, Alexander R. Gaudio, and Eliot L. Berson "The Association between Visual Acuity and Central Retinal Thickness Retinitis Pigmentosa," *IOVS*, Vol.46, No.9, pp. 3349-3354, 2005.
- [3] 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.
- [4] A Oishi, A Otani, M Sasahara, H Kojima, H Nakamura, M Kurimoto, 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.
- [5] 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.
- [6] 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.
- [7] Ki-Kwang Lee and Chang Hee Han, "Medical Diagnosis Problem Solving Based on the Combination of Genetic Algorithms and Local Adaptive Operations," *지능정보연구*, 제14권, 제2호, pp. 193-206, 2008.
- [8] D. Nasien, S.S. Yuhaniz, and H. Haron, "Statistical Learning Theory and Support Vector Machines," *Proceedings of the 2010 Second International Conference on Computer Research and Development, IEEE Computer Society*, pp. 760-764, 2010.
- [9] 강선경, 소인미, 김영운, 이상철, 정성태, "SVM 과 LDA를 이용한 마커 검출 및 인식의 성능 향상," *멀티미디어학회논문지*, 제10권, 제7호, pp. 923-933, 2007.
- [10] Vladimir N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Hoboken, 1998.
- [11] Vladimir N. Vapnik, "An Overview of Statistical Learning theory," *IEEE Trans. on Neural Networks*, Vol.10, No.5, pp. 988-999, 1999.
- [12] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification(2nd ed.)*, Wiley Interscience, New York, 2001.
- [13] 오일석, 패턴인식, 교보문고, 서울, 2008.
- [14] Chin-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin, A Practical Guide to Support Vector Classification, <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>, 2010.
- [15] D. Meyer, F. Leisch, and K. Hornik, "The Support Vector Machine Under Test," *Neurocomputing*, Vol. 55, Issues. 1-2, pp. 169-186, 2003.
- [16] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a Library for Support Vector Machines, <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2010.
- [17] Chih-Wei Hsu and Chih-Jen Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Trans. on Neural Networks*, Vol.13, No.2, pp. 415-425, 2002.



김 현 미

1994년 덕성여자대학교 전자계산학과 이학사
2001년 서강대학교 정보시스템전공 공학석사
2007년 창원대학교 컴퓨터공학과 박사수료
1994~2000년 성우그룹 성우전자, 텔파이코리아 전산실

2002~2004년 창원문성대학 컴퓨터정보처리과 연구교수
2005년~현재 창원대학교, 마산대학 등 강사
관심분야: 기계학습 및 패턴인식, 데이터마이닝



우 용 태

1982년 경북대학교 전자공학과 공학사
1984년 경북대학교 전자공학과 공학석사
1995년 경북대학교 전자공학과 공학박사
1987년~현재 창원대학교 컴퓨터공학과 교수

관심분야: 데이터마이닝, 온톨로지마이닝, 시맨틱 웹



정 성 환

1979년 경북대학교 전자공학과(공학사)
1983년 경북대학교 대학원(공학석사-정보통신)
1988년 경북대학교 대학원(공학박사-영상처리)
1983년~1985년 한국전자통신연구원 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