핵형 분류를 위한 패턴 분류기 구현

엄 상 희, 남 기 곤, 장 용 훈*, 이 권 순**, 정형환**, 김 금 석***, 전 계 록****
부산대학교 전자공학과, *동주여자전문대학 전산정보처리과, **동아대학교 전기공학과,
***부산대학교 컴퓨터공학과, ****부산대학교 병원 의공학과

The Implementation of Pattern Classifier for Karyotype Classification

S. H. Eom, K. G. Nam, Y. H. Chang*, K. S. Lee**, H. H. Chong**, G. S. Kim***, G. R. Jun****
Dept. of Electronic Engineering, Pusan National University,
* Dept. of Computer Information and Processing, Dong-Ju Women's Junior College,
** Dept. of Electrical Engineering, Dong-A University,
*** Dept. of Computer Engineering, Pusan National University,
**** Dept. of Medical Engineering, Pusan National University Hospital

Abstract

The human chromosome analysis is widely used to diagnose genetic disease and various congenital anomalies. Many researches on automated chromosome karyotype analysis has been carried out, some of which produced commercial systems. However, there still remains much room for improving the accuracy of chromosome classification.

In this paper, We propose an optimal pattern classifier by neural network to improve the accuracy of chromosome classification. The proposed pattern classifier was built up of multi-step multi-layer neural network(MMANN). We reconstructed chromosome image to improve the chromosome classification accuracy and extracted three morphological features parameters such as centromeric index(C.I.), relative length ratio(R.L.), and relative area ratio(R.A.). This Parameters employed as input in neural network by preprocessing twenty human chromosome images.
The experiment results show that the chromosome classification error is reduced much more than that of the other classification methods.

서 론

염색체(chromosome)의 세포유전학적(cytogenetics)인 해석은 인간의 유전학적 진단(diagnosis) 및 동·식물에 관한 유전학적 연구를 위하여 널리 사용되고 있다. 염색체의 분석은 핵형 분석(karyotype analysis)을 통하여 이루어지고 있으며 임상 의학 및 세포 유전학적 연구 분야 등에서 다양한 선천적 염색체성 유전질환 및 백혈병, 악성 종양의 진단 그리고 생물학적 연구 등을 위하여 매우 중요한 실험으로 알려져 있다. 특히 태아의 핵형 분석은 태아의 염색체의 이상을 조기에 판별하는 산전 유전 진단을 위한 필수적인 검사방법이다[1].

컴퓨터를 이용한 염색체 자동 분류에 관한 연구가 1964년 Ledly에 의하여 최초로 수행되어 임상에서 사용할 수 있음을 제시하였다. 1989년 Piper등은 염색체 영상에서 28개의 형태학적 특징 파라메타를 추출하여 패턴 분류를 위하여 통계적 패턴 분류 방법인 maximum likelihood방법을 사용하여 염색체를 자동 분류하는 연구를 수행하였으며, 1990년 Erik 등은 총 38개의 염색체 밴드 패턴정보를 6단계로 code화하여 구문론적 패턴 분류방법인 markov network방법을 사용하여 염색체를 분류하는 연구를 수행하였다[2]-[7]. 최근에는 인공 신경회로망(artificial neural network)을 이용하여 염색체를 분류하는 연구방법들이 제안되었다. 1994년 Lerner 등은 염색체 영상에서 농도 파일(density profile)에서 64개, 동원체 지수 및 염색체 길이 등 총 66개의 특징 파라메타를 추출하여 추출한 특징파라메타를 64d.p., 64d.p.+c.i. 및 64d.p.+c.i.+long.의 3가지로 입력패턴을 구성하여 two-layer feedforward neural network을 사용하여 패턴분류기를 구성하는 연구를 수행하였으며, 추출한 특징파라메타의 약 70%이상을 사용하여야만 염색체의 분류수행이 가능하다고 발표하였다[8][9].

패턴을 인식하고 분류하는 과정에서 원시영상(raw image)의 특징 정보를 소실하지 않고 특징 파라메타를 추출할 수 있는 영상 전처리(pre-processing) 기법이 아주 중요하다. 그러나 염색체 영상은 형태적으로 굽어진 염색체가 많이 존재하는 관계로 인하여 같은 번호의 염색체일지라도 동일한 특징 파라메타의 추출이 용이하지 않다. 또한 영상 전처리에 소요되는 시간이 상당히 길다. 따라서 영상정보의 손실이 적으면서 표본화된 특징 파라메타를 추출할 수 있는 영상 전처리 기법이 필요하다. 본 연구에서는 장용훈 등[10]이 제시한 염색체 영상의 재구성 기법(reconstruction algorithm)을 사용하여 염색체의 특징 파라메타를 추출하였다.

최근 패턴 인식 분야에 많이 사용되는 인공 신경회로망으로 염색체를 분류하는 연구에서는 하나의 다층 신경회로망(multi-layer neural network)구조를 사용하여 학습과 분류에 사용하고 있다. 또한 분류의 정확도를 높이기 위하여 많은 입력 파라메타들이 사용되고 있다. 그러나 다층 신경 회로망에서 입력 파라메타의 수가 증가하면 학습시간이

아주 많이 소요되며, 이런 파라메타들은 학습을 방해하는 요인으로 작용한다. 이러한 문제점을 해결하기 위하여 본 연구에서는 염색체 핵형 분류를 위한 패턴 분류기(pattern classifier)로 다단 다층 인공 신경회로망(multi-step multi -layer neural network : MMANN)을 제안한다. 제안된 MMANN은 염색체의 핵형분류도에 나타나는 개개의 염색체들이 형태학적으로 다른 패턴을 가지므로, 유사한 패턴을 가지는 부류들을 모아 염색체 군을 형성하고 있는 데에 착안하여 두단계의 다층 신경회로망을 구성하였다. MMANN에서는 염색체 특징 파라메타들을 1차적인 분류를 통하여 비선형성을 줄일 수 있으므로 분류의 정확도를 높일 수 있으며, 개별 염색체 군들에 따라 다층 신경회로망을 구분하여 학습시킬 수 있으므로 학습 소요시간을 감소시킬 수 있고, 또한 두단계로 구성된 8개의 다층 신경회로망을 사용하므로 최소의 특징 파라메타로 염색체를 분류할 수 있다.

본 연구를 위하여 임상적으로 정상인 20명의 염색체 영상을 재구성하여 동원체 지수(centromeric index:C.I.), 상대 길이비(relative length:R.L.)및 상대 면적비(relative area:R.A.) 등의 특징 파라메타를 추출하였다. 추출된 특징 파라메타들 중 10명분은 입력패턴으로 하여 다단 다층 인공신경회로망의 학습에 사용하였고, 나머지 10명분은 MMANN으로 구성한 패턴분류기의 분류 대상 입력패턴으로 사용하여 실험결과를 타 연구자들과 비교하였다.

재구성에 의한 형태학적 특징 파라메타 추출

염색체의 관찰은 유사분열(mitosis) 중기(metaphase)의 세포에서 가장 용이하다. 따라서 유사 분열 중기의 세포를 현미경을 통하여 관찰한 후, 이를 촬영한 필름을 스캐너(scanner)를 사용하여 임상적인 관점에서 정상인으로 판명된 20명 즉, 920개의 염색체를 촬영하여 영상파일을 저장하였다.

패턴 인식을 위하여 영상의 특징정보를 수치화 하여 그 영상의 고유성분을 추출하는 것을 특징 파라메타의 추출이라고 한다. 영상정보를 수치화 하여 입력 성분으로 선택하기 위하여 패턴 공간(pattern spaces)에서 군집화된 패턴 클래스들은 분류 정확도의 향상에 기여를 한다. 따라서, 적절하고 군집화된 특징 파라메타의 선택이 인식율을 향상할 수 있으며, 많은 양의 특징 파라메타들을 사용하여 나타날 수 있는 복잡도(complexity)의 감소를 위하여 영상의 특징을 가장 잘 나타내는 특징정보를 선택하여야만 한다. 염색체 영상의 형태학적인 특징 파라메타의 계산 방법은 표 1과 같다.

표 1. 특징파라메타 계산방법.
Table 1. The feature parameter calculation equation.

특징 파라메타	계산 방법
동원체 지수 (centromeric index : C.I.)	$C.I = \frac{\text{염색체 단완의 길이}}{\text{염색체의 길이}} = \frac{\ell_s}{\ell_c} \le 0.5$
상대 길이비 (relative length ratio : R.L.)	$C.I = \frac{\text{염색체 단완의 길이}}{\text{염색체의 길이}} = \frac{\ell_s}{\ell_c} \le 0.5$
상대 면적비 (relative area ratio : R.A.)	$C.I = \frac{\text{염색체 단완의 길이}}{\text{염색체의 길이}} = \frac{\ell_s}{\ell_c} \le 0.5$

본 연구에서 사용하기 위한 특징 파라메타는 염색체의 형태들이 비선형적인 구조를 가지고 있어 같은 염색체 번호에서도 동일한 특징 파라메타의 추출이 용이하지 않다. 따라서 염색체 영상의 중앙축의 화소를 기준으로 하여 32방사 방향으로 염색체의 폭을 구하여 선형적인 형태의 염색체로 재구성하여 특징 파라메타를 추출하였으며, 영상을 재구성하는 알고리즘의 순서도는 그림 1과 같다[10].

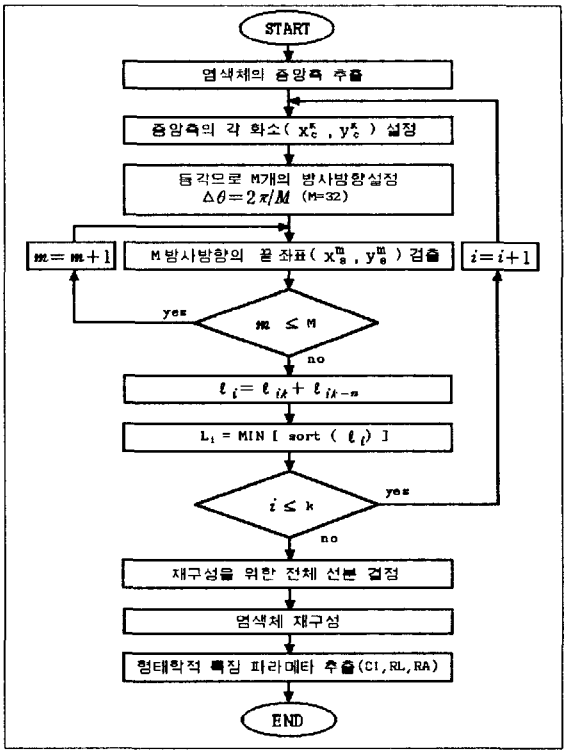

그림 1. 염색체 영상의 재구성을 위한 순서도.
Fig. 1. flowchart for chromosome image reconstruction.

그림 2는 재구성 알고리즘을 사용하여 1번 염색체 영상에서 재구성된 염색체 영상과 추출된 특징 파라메타의 결과를 나타내었다.

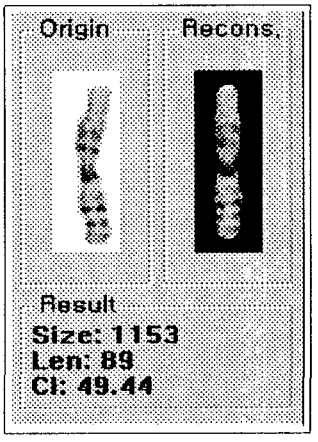

그림 2. 염색체 영상의 재구성 및 특징 파라메타.
Fig 2. The reconstruction of chromosome image and feature parameters.

신경회로망을 이용한 패턴 분류기 구성

신경회로망은 상호 연결된 뉴런에 의하여 입력과 출력사이에 비선형 사상(nonlinear mapping)을 하며, 이러한 사상 특성에 의해 변환, 필터, 최적화, 연상기억 및 분류 등의 기능을 수행할 수 있다.

염색체들은 형태학적으로 유사한 패턴을 가지는 부류들을 모아 염색체 군을 형성하고 있으며, 이를 기준으로 핵형분류도(karyogram)를 작성하여 염색체의 분류 및 진단에 활용하고 있다. 따라서 이러한 특징을 근거로 하여 염색체를 인식하고 분류하기 위하여 2단 다층 신경회로망(two-step MANN)을 사용한다. 첫 번째 단계의 다층 신경회로망(MMANN1)은 염색체를 7개 군으로 분류하기 위하여 구성하였고, 두 번째 단계의 다층 신경회로망(MMANN2)은 각 군내에 속한 개개의 염색체(1-22, X, Y)를 분류하기 위하여 7개의 다층 신경회로망으로 구성하였다. MMANN의 학습과 분류에 사용된 입력 파라메타들은 2절에서 구한 3가지의 특징 파라메타들을 사용한다. 표 2는 A에서 G까지인 7개의 염색체 군과 각 군에 포함되는 염색체 번호를 나타내었다.

표 2. 염색체 군과 염색체 번호.
Table 2. Chromosome group and number.

Chromosome Group	Chromosome Number	Chromosome Group	Chromosome Number
Group A	1번 ~ 3번	Group E	16번 ~ 18번
Group B	4번 ~ 5번	Group F	19번 ~ 20번
Group C	6번 ~ 12번, X	Group G	21번 ~ 22번, Y
Group D	13번 ~ 15번		

본 연구에서 제안한 MMANN은 그림 3과 같이 구성하였다. MMANN1은 입력층 뉴런 3개, 은닉층 뉴런 30개 및 출력층 뉴런 7개로 구성되며, 입력패턴은 C.I., R.L., 및 R.A.이며, 출력변수는 7개의 염색체 군(A~G)을 나타낸다. MMANN1에 의해 분류된 염색체 군의 출력은 MMANN2에 구성되어 있는 7개의 다층 신경회로망 중에서 해당되는 신경회로망을 선택한다.

MMANN2는 개개의 군에 해당되는 각각의 염색체를 분류하는 7개의 다층 신경회로망으로 구성되었고, 입력층 뉴런은 모두 동일하게 3개이며 출력층 뉴런의 수는 표 2에 나타낸 염색체 번호의 개수와 같다.

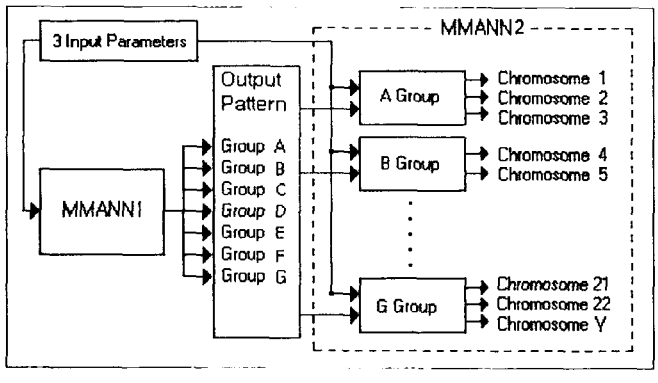

그림 3. MMANN의 구조.
Fig. 3. The MMANN structure for chromosome classification.

신경회로망에서는 가중치의 초기값, 활성함수의 기울기(slope of activation function), 학습 상수(learning constant), 중간층 뉴런의 개수 등이 학습에 큰 영향을 미치게 된다. 따라서 염색체의 특징벡터를 입력으로 하여 학습되는 신경회로망의 최적 상태를 정확히 정의할 수 없으므로 다양한 형태의 실험이 요구된다. 본 연구에서는 지도학습인 오차 역전파 학습알고리즘(error back-propagation learning algorithm)을 사용하였으며, 학습요소들의 적절한 값과 중간층 뉴런의 적합한 수를 선정하기 위하여 여러 번의 실험을 수행한 결과 표 3과 같은 파라메타를 얻을 수 있었다. 이때 사용된 교사신호는 분류와 일치하는 뉴런의 교사신호는 1.0이며 나머지는 0.0이다. 또한 학습 시간을 단축하고 신경회로망이 국부적인 극소점(local minimum)에 빠지지 않게 하기 위하여 모멘트법과 적응 학습법을 사용하였다.

표 3. 신경회로망의 파라메타.
Table 3. Parameters of neural network.

신경회로망 / 파라메타	MMANN1	Group A	Group B	Group C
활성함수의 기울기	1.0	1.2	1.1	1.0
학습 상수	0.05	0.05	0.01	0.05
중간층 뉴런의 수	30	10	10	30
신경회로망 / 파라메타	Group D	Group E	Group F	Group G
활성함수의 기울기	0.9	0.8	1.0	1.0
학습 상수	0.1	0.5	0.05	0.1
중간층 뉴런의 수	15	15	10	10

실험결과 및 고찰

핵형 분류를 위한 S/W환경은 Photoshop과 C 언어로 구현하였으며, 패턴 분류기로 구현된 신경회로망은 각 층의 뉴런수, 활성함수의 기울기, 학습 상수, 가속 상수, 입력패턴 수, 학습 회수 및 최대 허용 오차 등의 파라메타들을 키보드를 통하여 입력하므로써 신경회로망의 학습에 유연하게 대처할 수 있도록 설계하였고, 또한 학습되는 과정을 가시화하기 위하여 오차의 변화를 그래픽 처리하였으며, 추출된 결과는 표로 구성하여 파일로 저장하였다.

그림 4는 재구성된 염색체의 A그룹을 분류하는 MMANN2에서의 학습하는 과정을 SSE(sum-square error)의 변화로 나타낸 것이다. SSE 그래프는 급속한 곡선으로 변화하면서 상당히 빠른 속도로 학습이 진행되었음을 알 수 있다. 학습회수는 4,980회이며, 나머지 경우는 표 3과 같이 각 염색체 군에 따라 적절한 최대 허용 오차를 선정하였다.

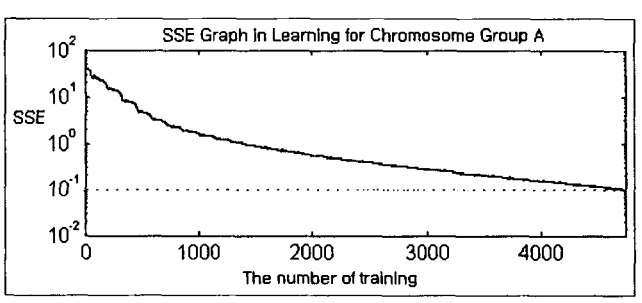

그림 5. MMANN2학습시의 자승 오차의 변화
Fig. 5. The variation SSE in MMANN2 learning

학습된 신경회로망을 패턴 분류기로 이용하여 정상으로 판명된 10명의 재구성된 염색체 특징 파라메타를 입력하여 염색체의 핵형 분류를 수행하였다. 표 4는 전체 분류 분류 대상 중 1명의 염색체를 분류한 결과를 나타내었다. * 표시는 오분류된 염색체를 나타낸다.

표 4. 염색체 샘플 1의 분류 결과.
Table 4. The classification results of chromosome sample 1.

No.	Chromosome Sample 1				No.	Chromosome Sample 1			
	No.	Group	No.	Group		No.	Group	No.	Group
1	1	A	1	A	13	13	D	13	D
2	2	A	2	A	14	*13* *	D	14	D
3	3	A	3	A	15	15	D	15	D
4	4	B	4	B	16	16	E	16	E
5	5	B	5	B	17	17	E	17	E
6	6	C	6	C	18	18	E	18	E
7	7	C	7	C	19	19	F	19	F
8	8	C	8	C	20	20	F	20	F
9	9	C	9	C	21	21	G	21	G
10	10	C	10	C	22	22	G	22	G
11	11	C	*8* *	C	X	X	C		
12	12	C	12	C	Y	Y	G		

분류대상인 10명의 염색체를 패턴분류기로 분류한 결과 염색체 군은 오분류 없이 정확히 100%를 분류하였다. 전체 분류 대상인 460개의 오분류 결과는 21개로 4.56%의 분류 오차율을 나타내었다. 각 군의 분류 오차는 표 5와 같으며, 표 6은 본 연구에서 수행한 방법과 타 연구 방법에 의한 분류오차의 결과를 비교하여 나타내었다.

표 5. 염색체 분류 오차.
Table 5. The error for chromosome classification.

	Group A	Group B	Group C	Group D	Group E	Group F	Group G	Total
오분류 개수	0	0	12	6	0	3	0	21
분류 오차율	0%	0%	2.61%	1.30%	0%	0.65%	0%	4.56%

표 6. 염색체 분류방법 및 결과.
Table 6. Classification methods and results comparison of chromosomes.

Researches	Methods	Classification error(%)	feature parameters	Remarks
Proposed	MMANN	4.56	3	
Lerner	Two-layer N.N.	언급 없음	66	특징파라메타의 약 70% 사용시 분류가능성 제시
Piper	Maximum likelihood	5.8 ~ 21	28	
Lucas	Non-parametric baye's rule	4.0 ~ 11.5	7	끝어진 염색체 분류대상 제외
Erik	Markov Network	7.9	38	

Lerner등은 농도파일(density profile)에서 64개, 동원체 지수 및 염색체 길이 등 66개의 특징 파라메타를 추출하였다. 그리고 추출한 특징파라메타를 64d.p., 64d.p.+c.i. 및 64d.p.+c.i.+long.의 3가지로 입력패턴을 구성하여 two-layer feedforward neural network을 사용하여 패턴분류기를 구성하는 연구를 수행하였으며, 추출한 특징파라메타의 약 70% 로서 최종적인 분류수행이 가능하다고 발표하였고, 분류오차에 대한 언급은 없었다. Piper 등에 의해 수행된 연구는 염색체의 면적, 상대농도, 길이, 동원체 지수 및 가중농도곡선 등 28개의 특징 파라메타를 사용하여 maximum likelihood방법으로 데이터 베이스를 구축하여 분류한 결과 영상의 질에 따라 5.8%~21%의 분류오차가 나타났다. Lucas등은 염색체 길이, 동원체 지수 및 5개의 농도정보 등 총 7개의 특징 파라메타를 사용하여 non-parametric baye's rule을 적용하여 영상의 질에 따라 분류오차가 4.0~11.5%로 우수한 결과를 나타내었지만, 형태학적으로 많이 구부러진 염색체는 실험 대상에서 제외한 단점을 가진다.

결 론

비선형적인 염색체 영상을 분류하기 위하여 입력 파라메타의 군집화를 위하여 영상 재구성 알고리즘을 사용하여 염색체 영상을 재구성하여 20명에 대하여 3종류의 형태학적인 특징 파라메타 C.I., R.L. 및 R.A.의 추출하였다. 그리고 패턴인식 분야에 많이 사용되고 있는 인공 신경회로망을 이용하여 핵형 분류에 적합한 패턴 분류기를 구성하기 위한 연구를 수행하였다. 실험 결과 염색체 군의 평균분류 오차는 없었으며, 전체 염색체에 대한 평균분류 오차는 4.56%를 나타내었다.

본 연구에서 제안한 MMANN은 학습시간을 최소화하여 염색체 패턴분류기를 구현하였으며, 염색체의 분류오차에서도 만족한 결과를 나타내었다. 또한 분류에 필요한 특징 파라메타의 수를 감소시키므로써 특징 파라메타를 추출하는 영상 처리 부분에서도 시간을 단축할 수 있어 염색체를 분류하는데 소요되는 전과정의 시간을 단축할 수 있었으며, 분류오차에서도 타 연구에 비하여 우수한 결과를 나타내었다. 분류 오차가 나타나는 염색체 군의 신경회로망에 다른 특징 파라메타를 부가하여 패턴 분류기를 개선하면 염색체의 분류오차를 더욱 감소시킬 수 있으리라 생각된다.

참 고 문 헌

[1] An International System for Human Cytogenetic Nomenclature(ISCN), KARGER, 1985.
[2] Robert S. Ledly, "High-speed automatic analysis of biomedical picture," Science, vol. 146, pp. 216-223, 1964.
[3] J. M. Cho, and D. H. Hong, "Computer- assisted karyotyping system of Giemsa-stained chromosomes (II)," Proc. of 1989 Korea-Japan Joint Conference on MBE, pp. 19-23, Sep. 21-22, 1989.
[4] Lucas J. van Vliet, Ian T Young, and Brian H. Mayall, "The athena semi-automated karyotyping system," Cytometry, vol. 11, pp. 51-58, 1990.
[5] Brian H. Mayall, James D. Tucker, Mari L. Christensen, Lucas J. van Vliet, and Ian T. Young, "Experience with the athena semi-automated karyotyping system," Cytometry, vol. 11, pp. 59-72, 1990.
[6] Jens Gregor and Erik Granum, "Finding chromosome centromeres using band pattern information," Comput. Biol. Med., vol. 21, No. 1/2, pp. 55-67, 1991.
[7] Jim Piper and Erik Granum, "On fully automatic features measurement for banded chromosome classification," Cytometry, vol. 10, pp. 242-255, 1989.
[8] Lerner B., Levinstein M., Rosenberg B., Guterman H., Dinstein I., and Romem Y., "Feature selection and chromosome classification using a multilayer perceptron neural network," IEEE International Conference on Neural Networks, vol. IV. 6/7, pp. 3540-3545, Jun. 28-Jul. 2, 1994.
[9] Key-Rok Jun, Sang-Hee Eom, Young-Hoon Chang, "Optimal neural network classifier for chromosome karyotype ciassification," AI Simulation, 96, Int. SCS, pp. 315-318, march, 1996.
[10] 장용훈, 이권순, 정형환, 엄상회, 최욱환, 전계록, "염색체 영상의 재구성에 의한 형태학적 특징 파라메타 추출," 대한의공학회지, Vol. 17, No. 4, pp. 545-552, 1996.