2003년도 대한전자공학회 하계종합학술대회 제26권 제1호

클러스터링을 이용한 C. elegans 행동표현형 분류

나 원, 백중환
한국항공대학교 전자,정보통신,컴퓨터공학부

Classification of C. elegans Behavioral Phenotypes Using Clustering

Won Nah, JoongHwan Baek
School of Electronics, Telecommunications and Computer Engineering
Hankuk Aviation University
E-mail : nahwon@mail.hangkong.ac.kr

Abstract

C. elegans often used to study of function of gene, but it is difficult for human observation to distinguish the mutants of C. elegans. To solve this problem, the system, which can be classified automatically using the computer vision, is studying now. In the previous works, they described the auto-tracking system and the egg-laying timing modeling, which are used to automated-classify system. In this paper, we use three kinds of features, which are related to movement, size and posture of the worm, and each feature is described mathematically and normalized. In experimental result, we validated the features for the hierarchical clustering. And we used the Calinski and Harabasz's method to find the appropriate cluster number.

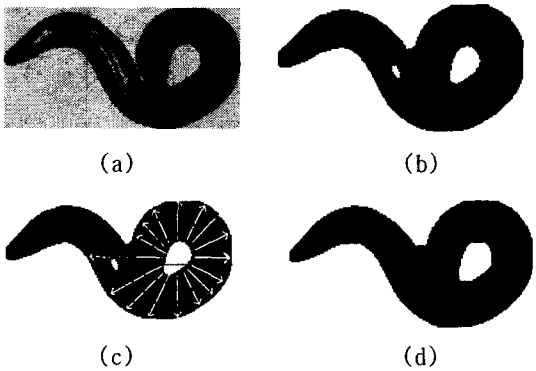
I. 서론

C. elegans 는 10^{-6}m 정도의 작은 선충이고 구조가 간단하여 신경과학분야에서 유전자 분석에 널리 사용된다. 선충의 전체 세포 수는 959 개로 이 중 302 개가 신경세포이며, 고등동물의 세포를 구성하는 분자적 구조물들 대부분이 존재한다고 한다. 유전자 변화와 C. elegans 의 움직임 사이의 관계를 통해 특정유전자의 역할과의 관계를 알아냄으로써 알츠하이머, 노화, 암 등의 질병 연구에 이용된다. 그러나 이러한 분류들은 모호하고 관찰자에 의해 주관적으로 분류되므로 같은 돌연변이가 다른 연구자에 의해 서로 다르게 기술되거나 명확히 구별 가능한 두 개의 돌연변이가 같은 종류로 분류되기도 한다. 이 문제를 극복하는 방법은 객관적인 분류가 가능한 자동화된 분석 시스템을 사용하는 것이다. 각 개체의 행동을 오랜 시간 기록하고 분석함으로써, 눈으로 구분하기 어려운 객관적인 분별과 돌연변이(Mutant)와 자연형(Wild-type)의 행동 편차를 정량적으로 분별하는 것이 가능해진다. 기존의 연구에서 자동화된 선충 추적 시스템을 제안하였고, 선충의 알을 낳는 시간을 통계적으로 모델링하였다[1][2].

본 논문에서는 C. elegans 선충의 영상을 이진화와 전처리를 통해 얻어진 세가지 분류의 특징값들을 이용한다. 이러한 특징값들을 이용하여 적절한 클러스터 수와 계층적 클러스터링의 결과를 얻는다. 실험에서 자동추적 시스템을 통해서 얻어진 영상 시퀀스를 이용하여 6 가지의 다른 선충 타입을 각각 100 개, 총 600 개의 샘플들을 이용하여 클러스터링 하였다.

II. 전처리와 특징값 추출

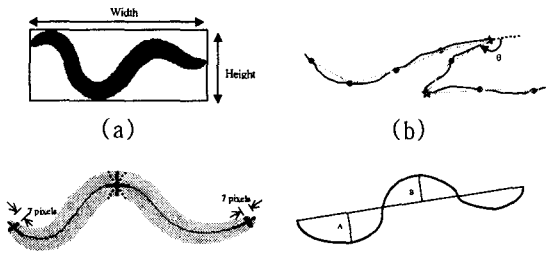
전처리 단계에서는 이진화, 미디언 필터링, 잡음과 홀(Hole)의 구분, 세선화 단계를 거친다. 먼저, 이진화 단계에서는 256 레벨의 그레이 이미지를 이진화한다. 본 논문에서는 선충의 몸통과 배경을 구별하는 방식으로 배경의 70%이하의 그레이 레벨값을 가지면 선충의 몸통으로 인식하고 그렇지 않으면 배경으로 인식하여 이진화를 진행한다. 그 후에 미디언 필터링을 수행하게 되는데 미디언 필터링을 통해서 임펄스성 잡음을 제거한다. 본 논문에서는 9x9 크기의 마스크를 이용하였다. 다음으로 잡음과 홀을 구별하는 단계를 거친다. 홀은 그림 1(a) 같은 선충의 몸통과 몸통이 닿으면서 생기는 구멍이다. 이를 위해 먼저 레이블링을 수행하고, 선충의 몸통과 배경을 제외한 레이블의 중심점을 구한다. 각각의 중심점에서 16 개의 동일한 각도 차를 갖는 벡터들을 이용하여 선충의 몸통인 부분에서의 거리를 측정한다. 각 벡터와 180 도를 이루는 벡터의 측정 거리를 합하고, 이렇게 더해진 8 개 벡터의 크기 중 최소값이 선충의 몸통의 두께보다 클 경우 홀로 판단하고 작으면 잡음으로 판단한다. 그림 1 에서 (b)는 이진화 결과에 미디언 필터링을 수행한 결과 영상이고, (c)는 홀 검색과정, (d)는 전처리 결과 영상이다.



(a) (b) (c) (d)

그림 1. 전처리 단계

다음으로 특징값 추출을 하게된다. 먼저 선충의 움직임에 관련된 특징들은 선충의 움직인 거리와 반전 빈도(Reversal Frequency)에 관한 특징값이 포함된다. 반전 빈도를 측정하기 위해서 선충의 반전을 인식해야 한다. 반전의 인식하는 방식은 중심점의 궤적을 통해서 움직인 각도를 알아내어 그 각(그림 2 (b)의 θ)이 120 도보다 클 경우 반전이 일어난 것으로 인식한다.



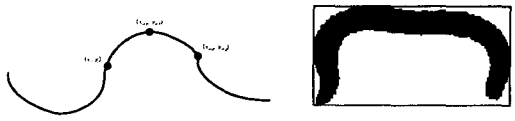
(a) (b) (c) (d)

그림 2. (a)MER, (b)움직임 궤적, (c)두께, (d)골격의 진폭

선충의 크기에 관련된 특징값들로 MER(Minimum Enclosing Rectangel)의 폭과 높이, 선충의 면적, 두께 등을 사용하였다. MER 은 그림 2 (a)와 같이 선충에 꼭 맞는 직사각형을 의미한다. 선충의 면적은 이진화된 선충의 몸통 픽셀의 개수를 카운트하고, 두께는 그림 2 의 (c)와 같이 선충의 양쪽 끝에서 7 픽셀과 중심에서 +5, 0, -5 도에서 각각 측정하여 최소값으로 결정한다.

선충의 모양에 관련된 특징값들로 선충이 고리모양을 만든 프레임의 수, 골격의 진폭, 각 변화율, Ω형태를 띤 프레임 수에 관한 특징값을 사용하였다. 고리모양의 형태는 선충의 몸통이 서로 닿으면서 생기는데 이것을 인식하기 위해서 레이블링의 결과 중 몸통과 배경을 제외한 다른 레이블이 발견될 때 인식할 수 있다. 선충의 진폭은 그림 2 (d)와 같은 선충의 골격에서 A(하향 진폭)와 B(상향 진폭)를 구한다. 각 변화율은 하나의 선충 골격위에 10 픽셀 간격의 점들을 정의했을 때 이러한 연속적인 점들의 각 변화의 평균이다.

각 변화율(R)은 아래의 수식을 이용하여 특징값을 구할 수 있다. 이를 통해서 선충의 꿈틀대는 정도를 알 수 있다.



(a) (b)

그림 3. (a)각 변화율, (b) Ω형태

$$R = \left\{ \frac{1}{n-1} \sum_{i=1}^{n-1} \theta_i \right\} / L \qquad (1)$$

$$\theta_i = \arctan\frac{y_{i+2}-y_{i+1}}{x_{i+2}-x_{i+1}} - \arctan\frac{y_{i+1}-y_i}{x_{i+1}-x_i} \quad (2)$$

Ω형태는 골격의 진폭을 이용하여 알 수 있고 아래 수식으로 정의된다. A_r 값이 0 인 경우 Ω형태로 결정한다.

$$A_r = \frac{\min(A,B)}{\max(A,B)} \quad (3)$$

위에서 언급한 여러 가지 특징값 각각의 최소, 평균, 최대값들을 이용하여 117 개의 특징값들을 얻었다. 각 특징값들은 0.0~1.0 사이의 값을 갖도록 수식 (4)를 이용하여 정규화 하였다. 여기에서 N 은 특징값들의 개수이다.

$$x_n = \frac{x_i - \min(x)}{\max(x)-\min(x)}, \quad i = 1,2,...,N \quad (4)$$

III. 계층적 클러스터링과 최적의 클러스터 개수

k-means 알고리즘은 초기값에 따라 결과가 바뀌는 불안정한 단점이 있기 때문에 본 논문에서는 계층적 클러스터링을 이용하여 선충을 분류하였다. 또한 최적의 클러스터링 결과를 얻기 위해 적절한 클러스터 수를 구하여 최종 분류 결과를 결정하였다.

3.1 계층적 클러스터링

본 논문에서는 계층적 클러스터링 방식 중 Ward[3]의 방식을 사용하였다. 이 방식은 초기에 각 샘플을 하나의 클러스터로써 시작하여 모든 클러스터쌍 사이에서 반복을 통해 가장 작은 자승오차를 가지는 쌍을 병합하여 새로운 클러스터를 만든다. 각 클러스터를 위한 자승오차는 다음과 같이 정의된다. 하나의 클러스터가 m 개의 샘플을 포함한다면, 샘플의 자승오차(Squared Euclidean Distance)는 다음과 같다.

$$\sum_{j=1}^{d}(x_{ij}-\mu_j)^2 \quad (5)$$

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \quad (6)$$

전체 클러스터들에 대한 자승오차(E)는 샘플들의 자승 오차들의 합이다.

$$E = \sum_{i=1}^{m}\sum_{j=1}^{d}(x_{ij}-\mu_j)^2 = m\sigma^2 \quad (7)$$

3.2 최적의 클러스터 개수

클러스터링은 클러스터 개수에 따라 결과에 큰 차이를 보인다. 그러므로 적절한 클러스터의 개수를 찾는 것은 클러스터링 방식을 사용함에 있어서 매우 중요한 일부분으로 생각할 수 있다. 본 논문에서는 최적의 클러스터 개수를 찾는 여러 가지 방식들을 비교했을 때 상대적으로 성능이 뛰어난 Calinski 와 Harabasz[4]의 방식을 이용하였다.

Calinski 와 Harabasz 방식에서 클러스터의 적절한 개수를 찾는데 VRC(Variance Ratio Criterion)라는 파라미터를 이용한다. N 개의 샘플을 k 개의 클러스터로 나누었을 때 VRC 은 수식 8 과 같다. WCSS(Within-Cluster Sum of Squares)는 클러스터 내부에서의 자승 오차이고, BCSS(Between-Cluster Sum of Squares)는 클러스터간 자승 오차를 나타낸다.

$$VRC = \frac{BCSS}{k-1} / \frac{WCSS}{n-k} \quad (8)$$

IV. 클러스터링 결과

본 논문에서는 6 개의 선충 돌연변이들(Wild, Goa-1, Nic-1, Unc-36, Unc-38, Egl-19)로 실험하였고, 각 돌연변이들은 100 개의 0.5 frame/sec 인 5 분 분량의 시퀀스를 사용하였다. 특징값들은 0.0 과 1.0 사이의 값을 갖도록 정규화 된 값들을 사용하였다.

그림 4 는 Calinski 와 Harabasz 방식을 이용하여 얻은 클러스터 개수에 따른 VRC 값을 나타낸다. 그래프에서 VRC 값은 감소하다 클러스터의 개수가 4 일 때 주변의 클러스터 개수가 3, 5 일 때의 값보다 큰 값을 가진다. 그러므로 최적의 클러스터 개수는 4 개로 결정된다.

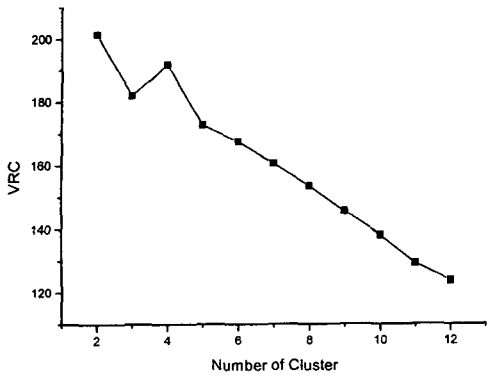


그림 4. 클러스터 개수에 따른 VRC 값

위에서 구한 최적의 클러스터 개수를 이용하여 클러스터 수가 4 개 일 때의 계층적 클러스터링 결과를 구하면 그림 5 와 같다.

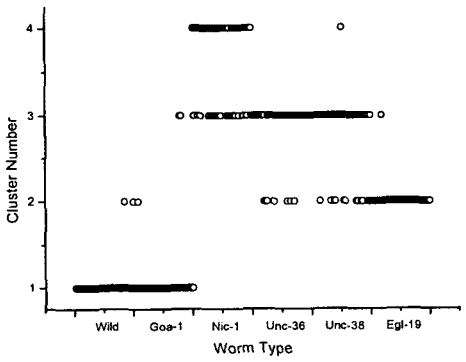


그림 5. 클러스터 수가 4 일 때 클러스터링 결과

표 1 을 살펴보면 클러스터 1 은 Wild 와 Goa-1 타입, 2 는 Egl-19 타입, 3 은 Unc-36 과 Unc-38 타입, 4 에서는 Nic-1 타입이 군집을 이루고 있음을 알 수 있다. 실제로 Wild 와 Goa-1, Unc-36 과 Unc-38 은 비슷한 습성을 보인다고 알려져 있으므로 클러스터링 결과에서도 그와 비슷한 성향을 띄는 것을 알 수 있다. 각 클러스터에 포함된 선충의 타입은 표 1 과 같다. 표에서 Nic-1 의 결과를 제외한 5 타입의 선충의 분류 성공율은 90%가 넘는 것을 알 수 있었다. 또한 Nic-1 과 Unc-36, Unc-38 선충 타입은 서로 비슷한 움직임, 크기, 모양을 보임을 알 수 있었다.

표 1. 계층적 클러스터링 결과

Cluster # / Worm Type	1	2	3	4
Wild	98	2	0	0
Goa-1	97	1	2	0
Nic-1	0	0	33	67
Unc-36	0	12	88	0
Unc-38	0	15	84	1
Egl-19	0	99	1	0

V. 결론

C. elegans 의 분류에 있어서 계층적 클러스터링을 이용하면 클러스터 개수에 따른 분류 성공율을 쉽게 알아볼 수 있다. 또한 초기값 선정에 따른 불안정한 결과가 나오지 않고 안정적인 분류결과를 알아 볼 수 있었다. 본 논문에서는 적절한 클러스터 개수를 알아보기 위해 Calinski 와 Harabasz 의 방식을 이용하였다.

실험 결과에서 적절한 클러스터의 개수는 4 개로 구해졌고, 계층적 클러스터링의 결과에서 클러스터의 수가 4 개일 때의 결과를 얻을 수 있었다.

클러스터링은 특징의 수에 관계없이 구분이 가능한 소수의 특징값들을 이용함으로써 더 좋은 결과를 얻을 수 있다. 또한 선충의 구별 가능한 특징값들을 추출한다면 더 좋은 결과를 얻을 수 있을 것이다.

참고문헌

[1] Waggoner, L., et al.. "Control of behavioral states by serotonin in Caenorhabditis elegans", *Neuron*, pp. 203-214, 1998.

[2] Zhou, G.T., Schafer, W.R., Schafer, R.W. "A three-state biological point process model and its parameter estimation", *IEEE Trans On Signal Processing*, pp. 2698-2707, 1998.

[3] Richard O. Duda, " Pattern Classification" , JOHN WILEY & Sons inc., 2001.

[4] Glenn W. Milligan., Martha C. Cooper. " An examination of procedures for determining the number of clusters in a data set" , *PSYCHOMETRICA*, Vol. 50, No 2, pp 159-179, 1985.