결정트리 분류기법 기반 유전자 계통수 추론†

김신석[0] 황부현
전남대학교
{sskim, bhhwang}@sunny.chonnam.ac.kr

Inference of Gene Phylogenetic Tree based on Decision Tree

Shin-Suck Kim[0] Bu-Hyun Hwang
Dept. of Computer Science, Chonnam National University

요 약

분자생물학의 급진적 발전은 현대 계통분류학에 큰 변혁을 가져왔다. 특히 유전의 근원물질인 DNA나 RNA를 분리·조작·분석하는 기술의 발전으로 이를 이용한 계통수 제작은 계통생물학의 중요한 실험방법으로 자리잡고 있다. 그 중 염기서열 비교 방법은 현재 유전자 계통수 제작에 가장 널리 이용되는 방법이다. 하지만 이러한 계통수는 각 객체간의 거리만을 표현하고, 객체군간의 차이는 설명하기 힘들다. 본 연구에서는 염기서열의 상대적인 특징(유사도)을 대신하는 염기서열의 총량과 염기 함량 등을 이용해 새로이 분류 기법 중 결정트리 방법에 적용하고, 종 분류의 유전적 모델을 설계한다. 또한 결정트리의 클래스인 종은 상위 클래스들을 포함하고 있어, 본 논문에서는 기존의 결정트리 분류자를 수정한 단계적 결정트리 분류자를 제안한다.

1. 서 론

생물정보학(Bioinformatics)이란 생물학 연구에 의하여 생성된 제반의 정보를 컴퓨터 및 전산 기법을 사용하여 저장, 검색, 분석, 가공하는 연구 분야이다. 현재 생물정보학에서 연구하는 기술들은 다음과 같다. 생물학 연구에 의해 생성된 정보의 저장과 검색을 위한 통합 데이터 베이스로서 데이터웨어 하우스 구축과 서열의 특성 및 진화적 관계를 파악하기 위한 서열 분석 알고리즘과 프로그램 등을 개발하고, 더불어 데이터 마이닝 기법을 통하여 새로운 지식들을 발견하고자하는 연구가 진행되고 있다[1].

마이닝 기법 중 데이터 분류(Classification) 기법은 이미 분류된 객체 집단군 즉, 학습 데이터(Training Data)에 대한 분석을 바탕으로 아직 분류되지 않는 객체의 소속 집단을 결정하는 작업이다. 현재까지 제안된 여러 가지 분류 모델(Classification Model) 중 결정트리(Decision Tree)는 인간이 이해하기 쉬운 형태를 갖고 있기 때문에 탐사적인 데이터 마이닝 작업에 특히 유용하다[2]. 본 논문에서는 서열 데이터 분석을 위해 서열 데이터의 총량과 염기 위치에 따른 총량 등의 값을 추출하고 이를 종 분류를 위한 속성으로 사용한다.

과거 유전자 계통수(Gene Phylogenetic Tree)들은 각 종간의 분류 특성에 대하여 설명하기 힘들었다. 이는 유전자 계통수가 단순히 각 객체의 염기서열의 차이(유사도)만을 가지고 계통수를 추론하기 때문이다. 본 논문에서는 각 종간의 분류 특성을 찾기 위하여 결정트리를 가지고 종 분류 특성을 모델링하며, 또한 종의 계층적인 특징을 표현하기 위해 생물 분류의 상위 단위에서 하위 단위로 분류하는 단계적인 분류자를 제안한다.

2. 관련 연구

2.1 서열 데이터

원시 데이터인 염기서열은 AGTC 네 개의 문자로 이루어진 문자열이다. 염기서열은 DNA → mRNA→ protein -----→ phenotype의 흐름을 갖는다. 이러한 일련의 유전자의 흐름에서 최종산물인 표현형(Phenotype)을 비교 관찰하기보다는 세포 내·외적 환경의 영향을 적게 받는 DNA 즉 서열데이터를 직접 비교 분석하는 일이 생물의 진화역사를 추적하는데 객관적인 중요한 단서를 제공할 수 있다. 최근 분자생물학의 발전으로 여러 종들의 유전체 염기서열 데이터가 누적되어 이들을 서로 비교할 수 있게 되었다[3].

2.2 유전자 계통수 추론

이 장에서는 현재 생물학에서 유전체 염기서열을 이용한 계통수 추론 도구의 하나인 MEGA[7]를 통해서 기존 유전자 계통수에 대하여 살펴본다.

유전자 계통수를 제작하는 대표적인 방법은 UPGMA (Unweighted Pair Group Methods using arithmetic Averages)방법과 절약분석을 이용한 방법 등이 있다. 이 중 UPGMA는 유사도를 바탕으로 하여, 유사도가 가장 높은 객체들을 우선적으로 하나의 군으로 모으는 방법이다. 2개 이상의 객체들로 이루어진 군간의 유사도는 각 객체의 유사도의 평균값을 취한다[3].

두 염기서열간의 유사도는 각 위치에서 염기가 같은가 다른가를 가지고 계산한다. 이러한 유사도 측정은 각 염기간의 변이 특성을 고려하여 계산해야 한다. 실제 생체 내에서는 A⇔G, C⇔G로 변환되는, 즉 같은 퓨린 염기가 퓨린 염기로, 피리미딘 염기가 피리미딘 염기로 치환되는 확률과 퓨린 염기가 피리미딘 염기로 치환될 확률이 다르기 때문이다. MEGA에서는 이러한 염기 변이의 특성 등을 고려하여 계통수를 추론한다. 다음 그림 1. 은 MEGA에서 유사도에 바탕을 둔 UPGMA방법으로 그려진 계통수의 예이다.

† 이 논문은 1999년도 한국과학재단 특정기초 연구비 지원에 의한 결과임(과제번호:1999-2-303-006-3)

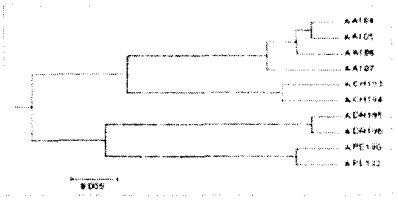


그림1 . MEGA 프로그램에서의 결과(UPGMA)

이렇게 생성된 계통수(그림 1.)에서 각 객체간의 거리는 알 수 있지만, 각 객체들 사이의 분류 속성은 알 수 없다. 또한 각 객체가 어떤 집단에 속하는지, 각 집단이 어떤 유전적 특성에 의해 결정되는지를 알 수 없기 때문에 새로운 자료의 집단 결정이 어렵다.

2.3 결정트리에 의한 분류

분류는 이미 분류된 데이터로부터 아직 분류되지 않은 객체의 소속 집단을 결정하는 것으로서, 이 중 결정트리를 사용하는 방법은 빠른 실행시간과 사람이 이해하기 쉬운 규칙을 생성한다. 그리고 결정트리 방법은 높은 차원의 데이터를 처리 할 수 있는 장점이 있다. 이러한 결정 트리 분류기법의 대부분은 다음 2단계로 이루어진다[4][5][6].

- 트리 성장 단계(Tree Building)
  각 속성을 평가하여 최적 분할 속성(best split attribute)을 선택하고, 이를 이용하여 학습데이터 집합을 최적 분할 속성값에 따라 부분집합으로 분할된다. 이러한 선택-분할 과정은 부분집합에 순환적으로 계속된다. 데이터 집합이 동일한 클래스이면 멈춘다.
- 트리 전지 단계(Tree Pruning)
  트리 성장 단계에서 완성된 트리에서 오류유발 데이터와 통계적 변동을 가지는 가지들을 제거하는 과정이다.

완성된 결정트리에는 각 클래스의 분류 모델이 나타나고, 이를 이용하여 분류되지 않는 새로운 객체의 클래스를 결정한다.

3. 단계적 결정트리 분류자

3.1 속성 추출

본 논문에서는 유전체 데이터를 염기서열에 나타나는 총량, 염기의 위치에 따른 가중치 합, 그리고 G+C 함유량의 3가지 특성으로 표현한다. 각 염기 아데닌(A), 구아닌(G), 시토신(C), 티민(T)에서 A와 G는 퓨린이라 하며, 피리미딘에 속하는 C 와 T 보다 더 큰 물질이다. 본 논문에서는 이를 바탕으로 (A,G)와 (C,T)에 각각에 서로 다른 값을 두어 그 합을 하나의 속성(염기의 총량:TW)으로 사용한다. 또한 DNA 염기서열의 각 A, G, T, C는 각각 차지하는 비율(특히 G+C의 함유량)과 각 염기위치에 따라 서로 다른 유전적 의미를 가지고 있다[3]. 본 논문에서는 이를 바탕으로 G+C 함유량과 각 염기 위치에 따른 가중치의 합(LocW)을 속성으로 사용한다.

본 논문에서 학습집합(Training set)의 클래스(분류목적 속성, 집단)로 사용되는 '종'은 그 상위 클래스들이 존재한다. 계통분류학에는 종(species)-속(genus)-과(family)-목(order)-강(class)-문(phylum)-계(kingdom)의 생물분류단계가 있다. 이러한 클래스의 계층적인 특성은 기존의 연구들에서 볼 수 없던 형태이며, 본 논문에서는 이를 다루기 위해 클래스 테이블을 두어 상위 클래스에서 하위 클래스로 분류해 간다.

3.2 속성테이블과 클래스테이블(Attribute Table and Class Table)

각 객체의 염기서열 데이터는 3.1절에서 설명한 TW, LocW, G+C의 속성과 '종'의 상위 클래스를 포함하는 클래스 테이블(표1. class table)로 변환된다. 표1.은 결정트리를 이용한 계통수 제작을 위한 학습데이터의 예이다.

표1. 속성 테이블 및 클래스 테이블

[속성테이블]

id	TW	LocW	G+C
1	250	1300	0.44
2	200	1200	0.48
3	200	1190	0.46
4	200	1250	0.47
5	220	1100	0.45
6	170	1050	0.51
7	180	1150	0.52
8	170	1040	0.53
9	180	1050	0.55

[클래스 테이블]

id	종	속	과
1	등줄쥐	붉은쥐	쥐
2	흰넙적	붉은쥐	쥐
3	흰넙적	붉은쥐	쥐
4	등줄쥐	붉은쥐	쥐
5	멧발쥐	멧발쥐	쥐
6	청서	다람쥐	다람쥐
7	다람쥐	다람쥐	다람쥐
8	청서	다람쥐	다람쥐
9	다람쥐	다람쥐	다람쥐

* 쥐과 MURIDAE 의 3가지 종과 *Apodemus agrarius coreae(붉은쥐속 등줄쥐), Apodemus speciosus peninsulae(붉은 쥐속 흰넙적다리 붉은쥐), Micromys minutus ussricus(멧발쥐속 멧발쥐),* 다람쥐과 SCIURIDAE 2가지 종 *Sciurus unlgalis coreae(다람쥐속 청서), Tamias sibiricus asiaticus(다람쥐속 다람쥐)*

3.3 분할 평가(Evaluating Splits)

결정트리 성장단계에는 한 속성에 의해서 분할되고 이를 평가하는 것을 포함하고 있다. 이는 한 속성에 대한 분할이 얼마나 잘 되었는지를 평가한다. 지금까지 많은 평가함수들이 제안되었다. 본 논문에서는 이들 중 Gini 지수[5]를 사용한다. 이는 데이터 집합의 순수함의 정도를 나타내는 척도인데, 어느 한쪽의 클래스로만 구성되어 있는 집합일수록 순수하다고 판단한다. 데이터 집합이 순수할수록 Gini지수는 작은 값을 갖는다. Gini 지수는 데이터 집합 T에서 $gini(T) = 1 - \Sigma p_j^2$로 표현된다. 여기에서 $p_j = \frac{j\ class의 빈도수}{전체\ example수}$ 이다.

본 논문에서는 한 분할 평가를 위해서 각 속성 이용하여 분할된 부분집합 Gini 지수의 평균을 이용한다.

3.4 정지 규칙

기존의 정지규칙은 분할된 부분집합이 동일한 클래스로만 구성되면, 선택-분할 과정을 멈추는 것이다. 본 논문의 단계적 결정트리 분류자는 계층적 클래스를 다루기 위하여 정지규칙을 다음과 같이 수정하였다.

- 수정된 정지규칙
  분할된 부분집합이 동일한 클래스 일 때 하위 단계의 클래스로 클래스가 표시되며, 더 이상 하위 단계의 클래스가 없으면 완료한다.

표1.에서 나타나는 최상위 클래스는 '과'가 되며, 분류단계는 '과' → '속' → '종'의 순서이다. 또한 기존에 분류자들은 최하위 클래스인 '종'에 대한 모델만을 생성할 수 있다. 반면 단계적 결정트리 분류자는 각 단계(과, 속, 종)별 모델이 생성된다. 이러한 하위 클래스의 모델 결정은 상위 클래스의 모델과 통합된다.

3.5 노드 분할 과정

본 절에서는 표1에서 제시된 예를 통하여, '과' →'속' →'종'의 순서를 이용하는 단계적인 결정트리의 성장 과정을 설명하고, 각 단계별 트리를 보여준다. 그리고 마지막으로 모든 클래스가 통합된 트리를 보여준다.

그림 2.는 5가지 종의 상위 클래스인 '쥐'과와 '다람쥐'과를 분류한 그림이다.

S_1

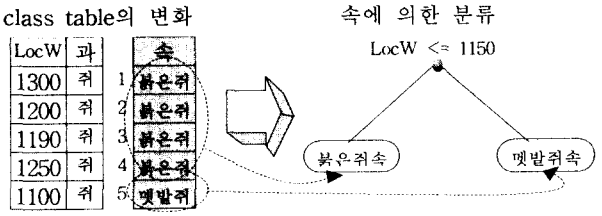
TW	LocW	G+C	과
250	1300	0.44	쥐
200	1200	0.48	쥐
200	1190	0.46	쥐
200	1250	0.47	쥐
220	1100	0.45	쥐

S_2

TW	LocW	G+C	과
170	1050	0.51	다람쥐
180	1150	0.52	다람쥐
170	1040	0.53	다람쥐
180	1050	0.55	다람쥐

그림 2. '과'(family)에 의한 분류

위 그림은 G+C ≤ 0.5가 쥐와 다람쥐를 구분하는 최적분할 속성임을 나타내고 있다. 그리고 최적 분할 속성으로 분할된 부분 집합은 S_1과 S_2 이다. 여기에서 S_1은 모두 동일한 클래스인 '쥐'를 나타내고 있기 때문에 수정된 정지 규칙에 의해 S_1은 쥐과의 하위 클래스인 붉은쥐속과 멧밭쥐속에 대하여 분류를 계속하게 된다(그림3).

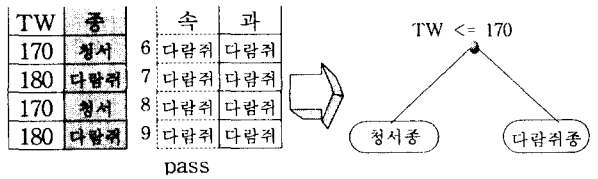


LocW	과
1300	쥐
1200	쥐
1190	쥐
1250	쥐
1100	쥐

	속
1	붉은쥐
2	붉은쥐
3	붉은쥐
4	붉은쥐
5	멧밭쥐

그림 3. S1(쥐 과)에 대한 속의 분류

위에서 대상 클래스는 '과'에서 '속'이 되고, '속'에 의해 찾아진 최적분할 속성은 LocW ≤ 1150임을 나타낸다. 그 결과 G+C ≤0.5이고 LocW≤1150이면 멧밭쥐속, G+C ≤ 0.5이고 LocW > 1150이면 붉은쥐속임을 나타낸다. 이러한 모델은 상위 클래스인 쥐과의 분류 속성과 통합하여 생성한다. S_2에서도 S_1과 같이 대상 클래스가 '속'이 된다. 하지만 S_2는 속 또한 모두 동일하기 때문에 다시 '속'의 하위 클래스인 종에 대해 분류한다(그림4).



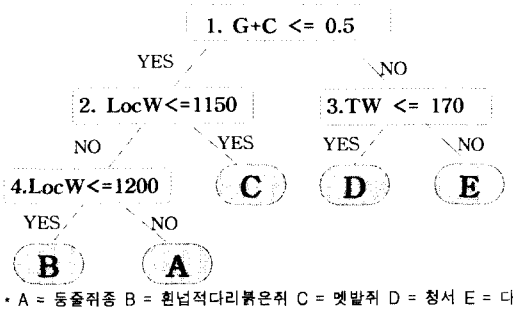
TW	종
170	청서
180	다람쥐
170	청서
180	다람쥐

	속	과
6	다람쥐	다람쥐
7	다람쥐	다람쥐
8	다람쥐	다람쥐
9	다람쥐	다람쥐

S2에서 모든 sample에 대하여 동일한 속을 나타낸다.

그림 4. S2의 종에 대한 분류

분할된 부분집합이 동일한 '종' 일 때, '종'의 하위 단계의 클래스가 존재하지 않으므로, 더 이상 분할하지 않는다. 위와 같은 과정을 통해서 생성된 최종 트리는 앞의 모든 단계가 통합되어 그림 5와 같이 그려진다.



* A = 동줄쥐종 B = 흰넓적다리붉은쥐 C = 멧밭쥐 D = 청서 E = 다람쥐

1. 다람쥐과와 쥐과에 대한 분류 속성
2. S_1(쥐과)의 속에 대한 분류 속성
3. 다람쥐과에 대한 분류 (종)
4. 쥐과의 붉은쥐속에 대한 종 분류

그림5. 최종 통합된 트리 모형

기존의 트리는 A, B, C, D, E에 대한 단일 결과만을 위한 트리를 생성하였다. 그러나 본 논문에서와 같은 각 단계별 모델은 생성하지 못한다. 또한 생물학적 객체의 특성은 한 종에 대한 상위클래스들이 존재하기 때문에 기존의 분류기법은 같은 클래스에 대한 다른 결과를 나타낼 수 있다.

4. 결론

본 논문은 결정트리에 의한 생물 객체의 종 분류 문제를 다루었다. 생물 객체는 계층적인 클래스를 가지고 있기 때문에 기존의 결정트리 분류 기법의 수정이 필요하다. 본 논문에서 제안한 단계적 결정트리 분류자는 생물의 계층적인 클래스 정보를 이용하여 종 분류에 효율적인 정보를 제공할 수 있다. 또한 '종'에 대한 모델 형성뿐만 아니라 상위 클래스인 '속'과 '과'에 대한 분류 모델을 생성한다. 이러한 특징은 기존의 계통수에서 찾을 수 없는 분류 속성을 예측한 것이다. 향후 유전자 데이터의 특성을 설명할 수 있는 다양한 속성들을 찾아 적용하고자 한다.

5. 관련연구

[1] 김정자, 이도헌, "서열분석을 위한 연관규칙 탐사," 한국 정보 과학회 봄 학술발표논문집 Vol.28. No.1

[2] 이도헌, "데이터 마이닝 : 개념 및 연구 동향," 데이터베이스 연구회지 13권 4호, 1996.

[3] 김기중, "분자생물학적 자료와 계통수 제작," 한국유전학회, 2000

[4] Manish Mehta, Rakesh Agrawal and Jorma Rissanen, "SLIQ:A Fast Scalable Classifier for Data Mining", EDBT 96, Avignon, France, March 1996

[5] L.Breiman et.al "Classification and Regression Trees," Wadsworth, Belmont, 1984

[6] J.Ross Quinlan "C4.5:Programs for Machine Learning," Morgan Kaufman, 1993

[7] MEGA : M. Nei(Pennsylvania State Univ.) group, IBM PC용 program