숨은 속성값기반의 련관규칙발굴과 그 응용

공혜옥, 허은향

련관분류(AC)는 련관규칙발굴과 분류기를 결합한 혼합수법으로서 속성값들사이의 관계를 밝혀내여 클라스를 배정함으로써 전통적인 분류방법들의 정확성을 더욱 높일수 있게 한다. AC에서 리용되는 련관의 대상들 즉 항목모임/속성값들은 주어진 자료기지의 업무(transaction)/레코드들에 명백히 표현되여있다.

순서형련관규칙[1]은 련관규칙에 포함되는 속성들사이에 음적으로 존재하는 순서관계를 론의하지만 그것도 역시 레코드들에 양적으로 존재하는 속성값들을 리용한다. 그러나 대용량 자료기지에서 자료들사이에 존재하는 련관관계는 레코드에 명백히 표현되지 않을수 있다.

프로모터와 같은 전사과정을 가능하게 하는 생물신호들을 인식하기 위하여 여러가지 기계학습방법들과 련관규칙을 리용하는 연구도 진행되였다.

프로모터령역인식문제는 DNA배렬을 형성하는 누클레오티드들사이의 련관관계에 기초하여 론의할수 있다. 이로부터 론문에서는 숨은 속성값기반의 련관규칙을 새롭게 정의하고 그것에 기초한 AC를 제기함으로써 정보론적관점에서 DNA배렬의 프로모터배렬을 이전의 방법들보다 훨씬 빠르게 예측하는 한가지 방법을 제기하였다.

$T=\{t_1, t_2, \cdots, t_n\}$ 을 자료기지 D 안의 실체(혹은 레코드)들의 모임, $A=\{a_1, a_2, \cdots, a_m\}$ 을 D 안의 속성들의 모임이라고 하면 매 실체는 A에 의하여 규정된다.

$\mathrm{val}(t_i, a_j)$ 를 실체 t_i에 대한 속성 a_j의 값이라고 하자.

D_j $(j=1, \cdots, m)$ 를 매 속성에 대한 값령역으로서 빈모임이 될수도 있다고 하자.

그러면 2개의 값령역의 직적 $D_{j_1}\times D_{j_2}$에서 정의되는 관계 r $(\le, =, \ge)$를 고찰할수 있다.

R를 $D_{j_1}\times D_{j_2}$에서 정의할수 있는 가능한 모든 관계들의 모임이라고 하자.

이제 우리는 자료기지안의 속성들사이의 여러가지 련관관계를 발견하기 위하여 숨은 속성값기반의 련관규칙(ARHVA)을 다음과 같이 정의한다.

정의 1 다음과 같은 지지도－믿음도조건들을 만족시키는 련관규칙

$$a_{j_1}, a_{j_2}, \cdots, a_{j_l} \to a_{j_1} r_1 a_{j_2} \cdots r_{l-1} a_{j_l}$$

을 숨은 속성값기반의 련관규칙이라고 한다. 여기서

$$\{a_{j_1}, a_{j_2}, \cdots, a_{j_l}\} \subseteq A,\ a_{j_u} \ne a_{j_s},\ u, s=1, \cdots, l,\ u \ne s,\ r_k \in R,\ k=1, \cdots, l-1$$

은 $D_{j_k}\times D_{j_{k+1}}$에서의 관계이고 D_{j_k}는 속성 a_{j_k}의 값령역이다.

① $a_{j_1}, a_{j_2}, \cdots, a_{j_l}$가 n개의 실체들중 $sprt$%에서 함께 발생(비지 않음)한다. 이러한 $sprt$를 규칙의 지지도라고 부른다.

② $T_r \subseteq T$는 $a_{j_1}, a_{j_2}, \cdots, a_{j_l}$들이 함께 발생하는 실체들의 모임으로서 $\forall t_{i_0} \in T_r$에 대하여 $\mathrm{val}(t_{i_0}, a_{j_1}) r_1 \mathrm{val}(t_{i_0}, a_{j_2}) \cdots r_{l-1} \mathrm{val}(t_{i_0}, a_{j_l})$이 성립된다. $conf = |T_r|/|T|$를 규칙의 믿음도라고 부른다.

정의 2 $\min sprt$, $\min conf$를 각각 최소지지도, 최소믿음도라고 할 때 $sprt \ge \min sprt$,

$conf \geq \min conf$ 인 숨은 속성값기반의 련관규칙을 흥미있는 련관규칙이라고 부른다.

먼저 프로모터배렬에 대한 련관분류기에 대하여 론의하자.

DNA배렬이 주어졌을 때 그것이 프로모터배렬을 포함하는가 포함하지 않는가를 예측하는 AC를 구축하는것은 숨은 속성값들사이의 관계 즉 DNA배렬의 물리화학적특성값들사이의 일정한 관계를 추출하여 진행할수 있다는데로부터 론문에서는 다음과 같은 절차를 리용하여 프로모터령역인식문제를 론의한다.

① 프로모터배렬을 포함하는 DNA배렬들의 모임(정의실체모임) *PT*와 프로모터배렬을 포함하지 않는 DNA배렬들의 모임(부의실체모임) *NT*를 준비한다.

② *PT*와 *NT*를 훈련자료로 하여 학습을 진행함으로써 미리 규정된 최소지지도와 최소믿음도를 만족시키는 모든 가능한 숨은 속성값기반의 흥미있는 련관규칙들의 모임 *PR*(정의실체)와 *NR*(부의실체)를 추출한다.

③ 분류대상으로 되는 DNA배렬이 주어지면 그것이 *PR*를 만족시키는 확률을 계산하여 0.5이상이면 정의실체로 분류(프로모터배렬을 포함)한다. 그밖의 경우에는 부의실체로 분류(프로모터배렬을 포함하지 않음)한다.

론문에서 고찰하는 훈련자료모임은 DNA배렬(A, C, G, T들의 렬)들의 모임으로서 일반적으로 매 배렬의 길이는 각이하다. 그러나 우리는 고정된 길이를 가진 DNA배렬들에 초점을 두고 고찰한다. 왜냐하면 동일한 기능을 수행하는 프로모터배렬들은 류사한 모찌브들이 조합되여 이루어지므로 길이를 고정시켜야 전사시작부위를 일치시켜 론의할수 있기때문이다. 그러므로 론문에서는 $t_i = a_{i1}a_{i2}\cdots a_{im},\ a_{ij} = A|C|G|T\ \ (i=1,\cdots,\ n;\ j=1,\cdots,\ m)$ 들로 이루어지는 *PT*와 *NT*를 훈련자료모임으로 준비한다. 즉 길이가 *m*인 DNA배렬들의 모임을 고찰한다. 여기서 모임의 원소개수 *n*은 훈련모임에 따라 다를수 있다.

*PT*와 *NT*로부터 정의실체들에 대한 *PR*와 부의실체들에 대한 *NR*를 얻기 위하여 DNA배렬의 구조적특징으로부터 임의의 길이의 련관규칙들이 아니라 길이가 2인 2원련관규칙들만을 발견하도록 한다. 2원련관규칙은 프로모터령역을 포함하는가 포함하지 않는가를 구분하는데 충분하다. 왜냐하면 2보다 큰 길이를 가진 련관규칙인 경우는 그것의 2원부분규칙들을 고찰하는것으로서 DNA배렬의 속성들을 충분히 평가할수 있기때문이다.

2원련관규칙들만을 탐색하므로 분류기의 훈련시간을 크게 단축할수 있다.

훈련단계에 들어가기 앞서 중요한것은 숨은 속성값기반의 련관규칙을 발견하기 위한 관계모임 $R=\{r_k\ \ (k=1,\cdots,\ l-1)\}$ 를 정의하는것이다. $r_k \in R\ \ (k=1,\cdots,\ l-1)$ 는 속성값들사이의 관계로서 이것은 DNA배렬(A, C, G, T들의 렬)로부터 2개의 누클레오티드들사이의 관계를 표현한다. 즉 2개의 누클레오티드들사이의 2원관계는 대응되는 물리화학적성질을 반영하는 수값들사이의 관계로 된다. 이러한 수값들은 누클레오티드들의 물리화학적성질들을 특징지을수 있는 계산값으로 표현된다.

관계모임이 주어지면 *PT*와 *NT*에 대하여 ARHVA발굴알고리듬을 적용하여 프로모터배렬을 포함하는 정의실체들을 분류하는 규칙들의 *PR*와 프로모터배렬을 포함하지 않는 부의실체들을 분류하는 규칙들의 모임 *NR*를 추출한다.

결국 *PR*와 *NR*가 프로모터배렬인식기의 역할을 수행하는 련관분류기이다.

훈련이 끝난 후에 새로운 실체 *S*(DNA배렬)를 *PR*와 *NR*에 적용하여 다음과 같은 추론을 진행한다.

① ARHVA발굴알고리듬을 리용하여 S에 대하여 PR를 만족시키는 ARHVA의 개수 N_{positive}를 결정한다. 따라서 S에 대하여 PR를 만족시키지 않는 ARHVA의 개수 $M_{\text{positive}} = =|PR|-N_{\text{positive}}$이다.

② ARHVA발굴알고리듬을 리용하여 S에 대하여 NR를 만족시키지 않는 ARHVA의 개수 N_{negative}를 결정한다. 따라서 S에 대하여 NR를 만족시키는 ARHVA의 개수 $M_{\text{negative}} =|NR|-N_{\text{negative}}$이다.

③ S가 정의실체로 분류되는 확률 $P_{\text{positive}} = (N_{\text{positive}} + N_{\text{negative}})/(|PR|+|NR|)$를 계산한다.

④ S가 부의실체로 분류되는 확률 $P_{\text{negative}} = (M_{\text{negative}} + M_{\text{positive}})/(|NR|+|PR|)$ 혹은 $P_{\text{negative}} = 1 - P_{\text{positive}}$를 계산한다.

$P_{\text{positive}} \geq P_{\text{negative}}$이면 S는 정의실체로 분류하며 그밖의 경우는 부의실체로 분류한다.

다음으로 우에서 정의한 ARHVA를 리용하여 논벼(*Oryza sativa japonica*)의 비생물학적 스트레스응답성유전자들의 프로모터배렬인식을 위한 실험적평가를 진행하자.

먼저 길이가 1 000bp인 논벼스트레스응답성유전자들의 프로모터들을 포함하거나(50개의 정의실체) 포함하지 않는(50개의 부의실체) 100개의 DNA배렬로 된 자료모임을 준비한다. 매 DNA배렬은 910bp의 상류 및 90bp의 하류배렬로 이루어진다. 여기서 DNA가 RNA로 전사되는 시작부위에 관하여 배렬의 시작위치는 −910이고 마감위치는 +90으로 한다. 결국 론문에서 고찰하는 훈련자료모임의 PT와 NT에 대하여 각각 n=50, m=1 000이다.

우리의 방법에서 2원련관규칙 즉 길이가 2인 련관규칙들은 DNA배렬의 2개의 속성(2개의 누클레오티드)들사이의 관계($\text{AC} \to \text{A}r_{k_1}\text{C}$, $\text{AG} \to \text{A}r_{k_2}\text{G}$, $\text{AT} \to \text{A}r_{k_3}\text{T}$ 등)이며 미리 규정된 최소지지도와 최소믿음도를 만족시킨다.

논벼스트레스응답성유전자들의 모임에 대한 관계모임 R를 규정하기 위하여 누클레오티드의 물리화학적성질로서 몰질량, 녹음점, 밀도, 중원소비, 위상극성표면적, 염기조성을 선정하였다.(표 1)

표 1. 누클레오티드들의 물리화학적성질을 반영한 수값

구분	특성이름	A	C	G	T
성질 1	몰질량	0.894 1	0.735 1	1.000 0	0.834 4
성질 2	녹음점	1.000 0	0.889 7	0.993 1	0.926 9
성질 3	밀도	0.727 2	0.704 5	1.000 0	0.559 0
성질 4	무거운 원소비	1.000 0	0.673 4	0.860 9	0.815 6
성질 5	위상극성표면적	0.836 7	0.701 6	1.000 0	0.604 9
성질 6	염기조성	0.968 1	1.000 0	0.997 6	0.967 3

표 1에서 매 성질의 누클레오티드에 대응하는 특성값들은 해당한 물리화학적성질의 수값들을 표준화한 값들이다. 실례로 성질 4의 무거운 원소비는 매 누클레오티드에서 C, N, O원소개수들의 비로서 A(6:4:0), C(4:3:1), G(5:5:1), T(5:2:2)를 표준화하여 얻은 값이며 성질 6의 염기조성은 GenBank자료기지에 있는 논벼(*Oryza sativa japonica*)의 전게놈에서 매 누클레오티드에 대한 염기조성을 계산하여 표준화한 값이다.

ARHVA를 적용하여 물리화학적성질의 특성값들에 기초한 4개의 누클레오티드들사이의 관계를 밝히는것이 목적이므로 표 1에서 성질 2, 4가 동등하며 마찬가지로 성질 3, 5가 동등하다는것을 알수 있다. 이로부터 동등한 성질들은 제외하고 성질 1, 6 즉 몰질량과 염기조성에 대응되는 특성값들사이의 호상관계(=, $<_1$, $<_6$, $>_1$, $>_6$)로부터 DNA배렬이 프로모터배렬을 포함하는가, 포함하지 않는가를 식별하는 2진분류를 진행할수 있다. 여기서 관계의 아래첨수는 표 1의 성질의 번호를 나타낸다. 2진분류에서 목표분류는 DNA배렬이 프로

모터배렬을 포함하는 경우는 1이고 포함하지 않는 경우는 0이다.

위수상관결수[2]의 절대값을 리용하여 누클레오티드속성들과 목표분류속성사이의 상관을 평가한 결과 성질 1, 6이 가장 높은 평균상관을 나타낸다는것을 밝혔다.

관계모임 R가 정의된 후 PT와 NT에 대한 전처리를 진행한다. 성질 1, 6으로부터 R가 정의되므로 그것들과 목표결과와의 상관들을 고려하여 일정한 턱값보다 작은 상관을 가지는 속성들은 무시하는 방법으로 전처리를 함으로써 분류의 정확성을 높이도록 하였다.

상관턱값은 속성들사이의 관계를 충분히 무시할수 있도록 최소상관에 가까운 값으로 정한다.

논벼스트레스응답성유전자들의 DNA배렬모임 PT와 NT에 대한 관계모임은 결국 몰질량과 염기조성의 특성값들의 크기관계로 정의된다. 이에 기초하여 ARHVA발굴알고리듬을 적용하기 위하여 우리는 $\min sprt$를 0.95로 정하고 $\min conf$는 0.75부터 0.05씩 감소시키면서 0.4까지 PR와 NR를 추출하여 프로모터배렬을 인식하는 련관분류기를 생성하였다.

표 2. 최소믿음도에 따르는 실험결과

최소 믿음도	분류 정확도	거짓 정의 실체수	거짓 부의 실체수
0.75	0.91	3	6
0.70	0.91	5	4
0.65	0.92	3	5
0.60	0.92	4	4
0.55	0.95	3	2
0.50	0.94	3	3
0.45	0.97	2	1
0.40	0.98	1	1

생성된 련관분류기를 리용하여 DNA배렬들에 대한 프로모터분류를 진행한 결과는 표 2와 같다.

표 2에서 분류정확도는 2진련관분류기를 리용한 논벼스트레스응답성유전자들의 DNA배렬모임에 대한 프로모터인식의 정확도이다.

표 2에서 보는바와 같이 믿음도가 작아지는데 따라 정확도는 높아지는 경향성을 보여주었다. 한편 분류속도는 믿음도가 작아지는데 따라 빨라졌으며 이것은 전통적인 탐색수법과 비교해볼 때 대단히 빠른 속도로 된다.

론문에서 제기한 방법은 대용량의 생물게놈자료기지에서 최신자료발굴기술을 리용하는 패턴인식방법으로서 전통적인 탐색방법들에 비하여 대단히 효과적인것으로 된다.

참 고 문 헌

[1] M. G. Aggarwal; International Journal of Advanced Research in Computer Science, 8, 9, 365, 2017.

[2] Pawel Cichosz; Data Mining Algorithms: Explained using R, Wiley, 321～324, 2015.

주체107(2018)년 12월 5일 원고접수

Association Rule Mining based on Hidden Values of Attributes and Its Application

Kong Hye Ok, Ho Un Hyang

We define a new association rule based on hidden values of attributes, construct association classifier, and provide a much fast and precise method to predict promoter regions in DNA sequences than the former ones on the view of informatics.

Key words: association classifier, association rule mining, promoter prediction