규칙 귀납법을 위한 알고리즘에 의한 진단 시스템의 성능 개선

현우석°
한국성서대학교 정보과학부
wshyun°@bible.ac.kr

A Performance Improvement of Diagnosis System using Algorithms for Rule Induction

Woo-Seok Hyun°
Dept. of Information and Science, Korean Bible University

요 약

기존의 규칙 귀납법(Rule Induction)은 양성적 추론(positive reasoning)과 음성적 추론(negative reasoning)을 잘 반영하지 못하고 있지만 의학 분야의 추론은 양성적 추론과 음성적 추론을 모두 포함하고 있다. 이것이 의학 전문가들이 귀납된 규칙을 해석하는데 어려움을 가지게 되며, 진단 과정을 위해서 규칙을 해석하는 것을 쉽게 진행할 수 없는 이유이기도 하다. 본 연구에서는 양성적 규칙들과 음성적 규칙들의 귀납법을 위한 두 가지 알고리즘을 적용한 진단 시스템인 DS-ARI(Diagnosis System using Algorithms for Rule Induction)을 제안한다. 제안하는 시스템과 기존 시스템을 비교해 보았을 때 제안하는 시스템에서 전문가의 지식을 보다 정확하게 표현하여 정확성을 높이게 되었다.

1. 서 론

규칙 귀납법(induction)은 결정적 규칙들(deterministic rules)의 귀납법과 확률적 규칙들(probabilistic rules)의 귀납법으로 구분된다[1-2]. 여기서 결정적 규칙들은 명제로 보여질 수 있는 if-then 규칙으로 나타나게 된다. 결정적 규칙의 조건부를 지원하는 예들의 집합은 C로 표시하며 D에 의해 표시되는 결론부에 속하는 예들의 집합의 부분집합이 된다. 즉 $C \subseteq D$ 라는 관계가 성립한다. 그러므로 데이터 집합(data set)에서 양성적 예들(positive examples)은 결정적 규칙들을 지원하게 된다.

반면에 확률적 규칙들은 확률정보를 가진 if-then 규칙들이다[2]. 집합 이론의 관점에서 C는 D의 부분집합이 아니고 유사하게 교차한다. 즉 $C \cap D \neq \varnothing$ 이고 $|C \cap D| / |C| \geq \delta$ 이다. 여기서 threshold δ는 교차집합의 유사 정도를 나타내는데 도메인 전문가(domain expert)에 의해서 주어진다. 그러므로 다수의 양성적 예들(positive examples)과 소수의 음성적 예들(negative examples)은 확률적 규칙들을 지원하게 된다. 결정적 규칙들과 확률적 규칙들의 공통점은 하나의 예가 그들의 조건부를 만족하게 된다면 양성적으로 그들의 결론을 귀납하게 된다는 것이다. 이러한 규칙들에 의한 추론을 양성적 추론(positive reasoning)이라고 부른다.

의료 전문가들은 양성적 추론(positive reasoning) 뿐만 아니라 음성적 추론(negative reasoning)도 사용하게 된다. 이들은 if-then 규칙의 결론들이 음성적 용어들(negative terms)을 포함하고 있는 것처럼 표현된 후보 규칙들 중에서 적절한 선택을 하기 위하여 음성적 추론을 사용하기도 된다. 예를 들면 환자가 두근거리는 고통(throbbing pain)은 아니지만 두통(headache)을 호소했을 때 편두통(migraine)이 높은 확률로 의심되어서는 안된다. 여기서 음성적 추론은 진단 과정(diagnosis process)의 탐색 공간(search space)을 감소시키는 데 중요한 역할을 한다[2]. 기존의 규칙 귀납법은 양성적 추론과 음성적 추론을 잘 반영하지 못하지만 의학 분야의 추론은 양성적 추론과 음성적 추론을 모두 포함하고 있다. 이것이 의학 전문가들이 귀납된 규칙을 해석하는데 어려움을 가지게 되며, 진단 과정을 위해서 규칙을 해석하는 것이 쉽게 진행될 수 없는 이유이기도 하다. 그러므로 음성적 규칙들은 전문가의 결정 과정들을 반영하는 규칙들을 귀납하기 위해서 뿐만 아니라 영역 전문가들이 해석하기 쉽도록 규칙들을 귀납하기 위해서 데이터베이스로부터 귀납되어 져야만 한다. 두 가지 추론은 모두 의학 전문가들과 컴퓨터들이 협력하여 사용되어지는 진단 과정을 향상시키기 위해서 매우 중요하다.

본 연구에서는 양성적 규칙들과 음성적 규칙들의 귀납법을 위한 두 가지 알고리즘[3]을 적용한 진단 시스템인 DS-ARI(Diagnosis System using Algorithms for Rule Induction)을 제안한다. 제안하는 시스템과 기존 시스템을 비교해 보았을 때 제안하는 시스템에서 전문가의 지식을 보다 정확하게 표현하여 정확성을 높이게 되었다.

2. 근접 기법(focusing mechanism)

의학 추론에서 중요한 특징 중의 하나는 여러 가지 가능성이 있는 후보들로부터 최종 진단을 선택하기 위해서 사용되어 지는 근접 기법(focusing mechanism)이다. 예를 들면, 환자가 두통을 호소했을 때 환자 병력, 물리적 검사 그리고 실험실 test 등을 기본으로 하여 60가지 이상의 질환을 고려하게 된다.

이러한 유형의 추론은 배타적 추론(exclusive reasoning)과 포괄적 추론(inclusive reasoning)으로 구성된다. 일반적으로 진단 과정은 다음과 같이 진행된다. 첫째, 배타적 추론은 환자가 특정 질환을 진단하기 위해서 필요한 증상을 가지지 않을 때 후보들로부터 그 질환을 배제시킨다. 둘째, 포괄적 추론은 환자가 특정 질환과 관련된 증상을 가지고 있을 때 배타적 처리과정(process)의 결과에서 해당하는 질환을 의심하게 된다. 이러한 두 가지 단계들은 양성적 규칙과 음성적 규칙을 사용하여 모델화되는데 , 전자는 배타적 추론에 대응되며 후자는 포괄적 추론에 대응된다.

3. 규칙 귀납법을 위한 알고리즘(Algorithms for Rule Induction)

음성적 규칙 혹은 배타적 규칙의 대치(contrapositive)는 그림 1에서와 같은 알고리즘[3]에 의해서 배타적 규칙으로 귀납되어 지며 다음과 같이 처리된다. 첫째, L로 표시되는 속성, 값 쌍의 목록으로부터 기술자(descriptor) $[a_i = v_j]$를 선택한다. 둘째, 이 기술자가 D에 의해 표시되는 긍정적 예들의 집합과 교차하는 지를 확인한다. 셋째, 이 기술자는 긍정적 규칙들을 위한 후보 목록에 포함되어 지고 알고리즘은 그것의 적용범위(coverage)가 1.0과 같은지 아닌 지를 확인한다. 만약에 적용범위가 1.0과 같다면, 이 기술자는 D의 배타적 규칙의 조건부를 위한 수식인 R_{er} 에 추가된다. 넷째, 이 때, $[a_i = v_j]$가 리스트 L로부터 삭제된다. 첫째부터 넷째까지의 처리과정이 L이 empty가 되지 않는 한 반복된다. 다섯째, 최종적으로 L이 empty가 되었을 때, 이 알고리즘은 귀납된 배타적 규칙들을 대치시켜서 음성적 규칙들을 생성한다.

양성적 규칙은 그림 2에서와 같은 알고리즘에 의해서 포괄적 규칙으로 귀납되어 지는데 양성적 규칙의 귀납을 위해서 정확도의 적용범위의 threshold는 각각 1.0과 0.0으로 정했다. 이 알고리즘은 다음과 같은 방식으로 처리된다. 첫째, 단지 하나의 기술자로 구성된 수식의 리스트를 나타내는 L_1을 그림 1에서 나타난 알고리즘에 의해서 생성된 리스트 L_{er}로 대치시킨다. 둘째, L_1이 empty가 될 때까지 다음 (2-2)에서 (2-2)까지의 단계들을 반복 수행한다.

(2-1) 수식 $[a_i = v_j]$를 L_1에서 제거한다.

(2-2) 알고리즘은 $\alpha_R(D)$가 threshold보다 큰지 아닌지를 확인한다. (양성적 규칙의 귀납법을 위해서 $\alpha_R(D)$가 1.0 과 같은지 아닌지를 확인하는 것과 같다.) 만약 그렇다면, 이 수식은 양성적 규칙의 조건부의 리스트에 포함되어 진다. 그렇지 않다면, 그 수식은 결합하기 위해서 사용되어 지는 M에 포함되어 진다.

셋째, L_1이 empty일 때 , 다음 리스트, L_2가 리스트 M으로부터 생성된다.

```
procedure Exclusive & Negative Rules;
var
L: List; /* A list of elementary attribute-value pairs */
begin
L:= P0 ;/* P0: A list of elementary attribute-value pairs given in a database */
while (L ≠ { } ) do
   begin
     Select one pair [ai = vj] from L;
     If ([ai = vj]A ∩D ≠ {}) then do /* D: posotive esamples of a target class
                                        d */
       begin
         Lir := Lir + [ai = vj]; /* Candidates for Positive Rules */
         If (K[ai = vj](D)=1.0)
             then Rer:= Rer [ai = vj];
                /* Include [ai = vj] into the formula of Exclusive Rule */
       end
     L := L - [ai = vj];
   end
   Construct Negative Rules: Take the contrapositive of Rer ;
end {Exclusive & Negative Rules};
```

그림 1 배타적이고 음성적 규칙들의 귀납법

```
procedure Positive Rules;
var
  i:integer; M, Li:List;
begin
Li:= Lir; /* Lir: A list of candidates generated by induction of exclusive
            rules*/
  i:=1; M:={};
for i:= 1 to n do /* n: Total number of attributes given in a database */
   begin
     while (Li ≠ {}) do
       begin
          Select one pair R=∧ [ai = vj] from Li;
            Li := Li - {R};
            if (αR(D)>δα) then do
                          Sir:= Sir + {R};
             /* Include R in a list of the Positive Rules */
              else M := M + {R};
       end
     Li+1 :=(A list of the whole combination of the conjunction formula in M)
   end
end {Positive Rules};
```

그림 2 양성적 규칙들의 귀납법

4. 성능 평가

시뮬레이션 환경 하에서 제안하는 시스템의 성능을 평가하기 위해서 양성적 규칙들과 음성적 규칙들의 귀납법을 위한 두 가지 알고리즘[3]을 적용한 DS-ARI(Diagnosis System using Algorithms for Rule Induction)을 개발하였다. DS-ARI는 표 1과 같이 두통과 수막염 영역에 적용되었다.

표 1 실험을 위한 데이터베이스

Domain	Sample Size	Classes	Attributes
Headache	5629	19	195
Meningitis	1211	5	51

성능을 비교하기 위해서 실험은 다음과 같은 단계로 수행되었다. 첫째, 표 1의 예들은 임의로 새로운 training sample들과 새로운 test sample들로 나눈다. 둘째, 기존의 규칙 귀납법인 AQ15[1]와 C4.5[4]가 규칙 생성을 위해 새로운 training sample들에 적용되었다. 셋째, 귀납된 규칙들과 전문가에 의해 손으로 획득된 규칙들이 새로운 test sample들에서 test되었다. 이 과정은 100번 반복되었으며 100번 시도한 것에 대한 평균 정확도(averaged accuracy)를 구하였다.

실험결과는 표 2와 같다. 첫 번째와 두 번째 열은 제안하는 DS-ARI를 사용해서 얻어진 결과를 보여준다. 첫 번째 열은 양성적 규칙들과 음성적 규칙들을 모두 사용했을 때 얻어진 결과이고, 두 번째 열은 양성적 규칙들만 사용했을 때 얻어진 결과이다. 세 번째와 네 번째 열은 C4.5와 AQ15에서 얻어진 결과를 나타내 준다. 다섯 번째 열은 의학 전문가로부터 얻어진 결과이다. 이 결과들은 양성적 규칙과 음성적 규칙을 함께 사용하는 것이 의학 전문가의 규칙보다는 좋지 못하지만, 양성적 규칙만 사용하는 것보다 정확도를 향상시킨 것을 보여준다.

표 2 실험 결과 (평균 정확도)

Method	Headache	Meningitis
DS-ARI (positive + Negative)	90.5%	91.5%
DS-ARI (Positive)	65.3%	73.8%
C4.5	83.7%	
AQ15	85.2%	81.4%
Experts	95.0%	93.7%

5. 결론 및 향후 과제

제안하는 양성적 규칙들과 음성적 규칙들의 귀납법을 위한 두 가지 알고리즘[3]을 적용한 DS-ARI에서는 전문가의 지식을 보다 정확하게 표현하여 정확성을 높이게 되었다. 이것은 기존의 규칙 귀납법이 양성적 추론과 음성적 추론을 잘 반영하지 못하였지만 제안하는 시스템에서는 양성적 추론과 음성적 추론을 모두 포함하고 있어서 의학적 지식을 보다 효율적이고 정확하게 표현할 수 있었기 때문이다.

향후 연구 과제로는 확장된 양성적 규칙들과 음성적 규칙들의 결합이 양성적이고 음성적인 결정적 규칙들에 비해 성능을 더욱 향상시키는 가에 대한 연구가 남아있다.

참고문헌

[1] Michalski RS, Mozetic I, Hong J, and Lavrac N, "The multi-purpose incremental learning system AQ15 and its testing application to three medical domains", Proc. 5th Nat. Conf. Artificial Intelligence, pp. 1041-1045, 1986.

[2]Tsumoto Sand Tanaka H, "Automated discovery of medical expert system rules from clinical databases based on rough sets", Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, pp. 63-69, 1996.

[3] Shusaku Tsumoto, "Automated discovery of positive and negative knowledge in clinical databases", IEEE engineering in medicine and biology, pp.56-62, July/August, 2000. .

[4] Quinlan JR, C4.5 -Programs for machine learing, Palo Alto, CA: Morgan Kaufmann, 1993.