지식기반 시스템의 규칙 경향 표현에 관한 연구

최락현, 손창식*
DGIST ICT연구본부, 지능형로봇연구부
e-mail: choimosi@dgist.ac.kr, changsikson@dgist.ac.kr

A Study on Rule Trend Representation for Knowledge-based System

Rock-Hyun Choi, Chang-Sik Son*
Division of Intelligent Robot, ICT Research Institute, DGIST

요 약

건강검진 데이터 공유로 일반인의 다양한 상태 분석 연구가 가능해졌다. 규칙추론 알고리즘으로 다양한 건강 데이터의 패턴을 파악할 수 있으나, 산출된 규칙이 많아지면 쉽게 해석하기가 힘들다. 본 연구에서는 스타차트를 활용하여 다수의 규칙을 보다 편하게 표현 할 수 있는 방법을 제안했다.

1. 서론

대량의 정보 데이터가 공유되고 있어 데이터 분석 관련 연구가 많이 진행되고 있다. 특히 건강검진 정보가 일부 무료로 공유 되면서 국민건강 상태 분포에 대한 연구가 가능하다. 그 중 BMI와 허리둘레에 따른 분류로 비만학회에서 제공하는 동반질환 위험도 분석은 일반인들에게 경각심을 일깨워 질병에 걸리기 전에 건강관리의 경각심을 일깨워 줄 수 있다. 본 논문은 이 전 연구[1]에서 다룬 동반질환 위험도 데이터에서 추론된 규칙 결과를 보다 효과적으로 분석할 수 있는 방법을 제안한다. 다수의 산출된 규칙의 전반적인 경향을 볼 수 있는 방안으로 보다 쉽게 규칙을 파악할 수 있다.

2. 본론

본 논문의 실험의 초반부는 이전 연구와 일치한다. 2016년 100만 건의 국민건강검진 데이터를 사용했고, 입력값으로 '수축기혈압', '이완기혈압', 'BMI', '식전혈당(공복혈당)', '총콜레스테롤', '감마지티피', '흡연유무'를 동반질환 위험도 분석의 변수로 사용했다[1]. Class label인 동반진단 위험도는 비만학회에서 제안됐고 6가지로 분류 한다. 1-낮음, 2-보통, 3-높음, 4-약간 높음, 5-매우 높음, 6-가장 높음과 같이 분류된다[2]. 유사한 신체 상태를 가진 사람들의 데이터 패턴을 보기위해 남성, 37세, 키 176센티, 몸무게 78킬로인 데이터로 필터 하였다. 이전 논문과 같이 이산화에 MAD, 규칙추론에 PRISM 알고리즘을 각각 사용했다. 데이터 필터에 따라 3363건의 데이터가 추출되고 규칙추론을 통해 440건의 규칙이 산출 된다[1].

본 논문에서 제안하는 방법은 그림 1에서와 같이 아래쪽 보다 검은 3가지 단계를 나타낸다. 먼저 Class label 분포를 보았을 때 규칙에 따라 산출된 Class label은 3, 4 였다. 여기서 본 논문에서 제안된 방법에 따라 첫째로 산출된 규칙을 통해 Class label 분포 경향을 보기 위해 Coverage 값이 제일 높은 규칙을 기준으로 정렬하였다. 그 후, 가장 빈도가 높은 규칙의 첫 번째 R_1 값(추론된 규칙의 첫 번째 컬럼)과 C_n 값의 관계를 차트(그림2)로 표현하였다. 두 번째는 R_1(생성된 규칙 중 하나의 특징 값 컬럼)에서 가장 빈도가 높은 특징 값과 R_2의 관계를 차트로 표현 했다(그림3).

(그림 1) 실험과정

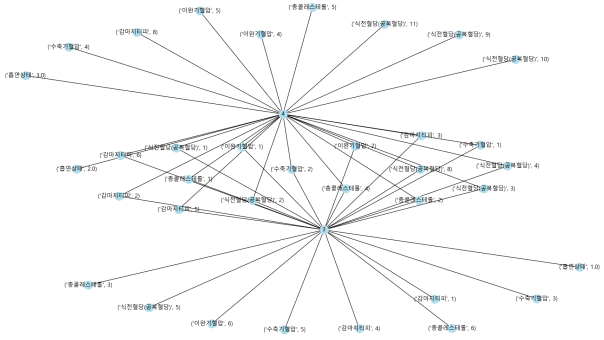
<표 1> MAD 이산화 결과

Feature	Cut point											
	1	2	3	4	5	6	7	8	9	10	11	
총콜레스테롤	94	199.5	248.5	271.5	289.5	328	933					
수축기혈압	88	120.5	130.5	150.5	163	173						
이완기혈압	51	80.5	88.5	101	103.5	107	130					
식전혈당	61	91.5	95.5	99.5	104.5	118.5	138.5	175.5	263.5	296.5	438	541
감마지티피	5	26.5	45.5	78.5	122.5	214.5	350.5	546.5	596			

그림 2를 통해 클래스 레이블에 따른 빈도가 가장 높은 규칙의 첫 번째 특징들의 전체 값과 Class label에 대해 3(높음 354건)과 4(약간 높음 86건)의 공통된 변수들을 제외하고 다른 특징을 보면 관계를 보면 다음과 같다.

- 클래스3: 비흡연자, 감마지티피 1 4, 식전혈당 5, 수축기혈압 3 5, 총콜레스테롤 3 6
- 클래스4: 흡연자, 감마지티피 8, 식전혈당 9-11, 수축기혈압 4, 총콜레스테롤5

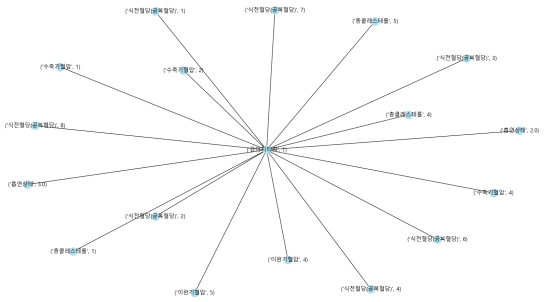
즉 1단계 비만에 해당되는 사람이 흡연상태, 감마지티피, 식전혈당만 비교해보면 모두 수치가 더 높은 경향을 나타내는 것을 확인할 수 있다.



(그림 2) 클래스 레이블에 따른 규칙 분포 경향

그림 3에서는 Class label 3일 때 빈도가 가장 높았던 감마지티피 1과 R_2와 관련된 변수를 살펴보았다. 식전혈당은 Class label 4 보다 다소 낮은 경향의 규칙 변수들이 분포 했다. 이완기혈압, 수축기혈압, 총콜레스테롤은 비교하기 힘든 분포를 보인다. 다만 흡연상태가 과거흡연자(1건), 흡연자(8건)으로 1단계 비만자와 비교했을 때 빈도가 낮지만 관계가 중복되는 경향을 보여준다.

- 식전혈당: 1, 2, 3, 4, 6, 7, 8
- 이완기혈압: 4, 5
- 수축기혈압: 1, 2, 4
- 총콜레스테롤: 1, 4, 5
- 흡연상태: 과거흡연자, 흡연자



(그림 3) 감마지티피와 관련된 규칙 분포 경향

3. 결론

PRISM으로 불리는 Covering 기법의 규칙추론 방법을 통해 추론된 규칙들의 경향을 표현해 보았다. 많은 수의 데이터와 다양한 분포를 가진 데이터 셋을 통한 분석에서는 산출되는 규칙의 수가 많기 때문에 전반적인 특징 파악을 위한 방법이 필요하다. 하나의 규칙 셋이 완벽하게 다른 특징 값들과 분류 될 수 없기 때문에 다소 중복되는 경향이 발견되지만, Class label에 따른 전반적인 규칙 경향을 Star chart로 표현하여 규칙의 경향을 보다 쉽게 파악할 수 있었다. 그러나 규칙집합의 빈도가 높은 변수로 표현을 했지만, 특정 클래스에서 하나의 규칙 집합 자체가 독립이기 때문에 하나의 규칙 변수로 관계를 표현할 수 없는 한계가 있다. 향후 연구에서 규칙집합에서 중복되는 규칙 특징 값을 처리하는 방법을 연구하여 $R_1 ... R_n$에 이르는 규칙집합의 경향과 중요도를 보다 쉽게 표현할 수 있는 방법에 대한 연구가 필요하다.

Acknowledgement

이 연구는 2019년도 산업통상자원부 및 한국산업기술평가관리원(KEIT)의 산업기술혁신사업 연구지원에 의한 연구입니다(10063553).

참고문헌

[1] 최락현, 이동하, and 손창식, “지식기반 복부비만 관련 동반질환 위험도 데이터 패턴분석,” 대한전자공학회 하계학술대회, 2019, pp. 1221 - 1222.

[2] M. H. Seo et al., “2018 Korean Society for the Study of Obesity Guideline for the Management of Obesity in Korea,” Journal of Obesity & Metabolic Syndrome, vol. 28, no. 1, pp. 40 - 45, Mar. 2019.