연관규칙 시각화를 위한 구조화된 연관맵

Structured Association Map for Visualizing Association Rules

김 준 우
동아대학교

Kim Jun Woo
Dong-A University

요약

연관규칙 탐사는 대표적인 데이터 마이닝 기법 중의 하나로, 트랜잭션 데이터에 포함된 항목들 간의 인과 관계를 의미하는 연관규칙의 추출을 목적으로 한다. 연관 규칙 탐사의 주된 문제 중 하나는 추출된 연관규칙의 수가 많을 경우, 이들을 적절히 해석하고 활용하는 것이 어렵다는 점이다. 이러한 문제를 해결하기 위해 본 논문은 구조화된 연관맵이라는 새로운 시각화 방법을 제안하고자 한다.

Ⅰ. 서론

연관규칙 탐사는 대표적인 데이터 마이닝 기법 중의 하나로, 트랜잭션 데이터로부터 유용한 연관규칙들을 추출하는 것을 목적으로 한다. 연관규칙은 일반적으로 $X \rightarrow Y$ 형태로 표현되며, 이 때 X, Y는 각각 트랜잭션 데이터를 구성하는 항목들의 집합을 의미한다. 특히, X를 전항(antecedent), Y를 후항(consequent)이라 지칭하고, 동일 트랜잭션 내에서 X와 Y의 인과관계가 높을수록 해당 연관규칙이 유용한 것으로 본다. 이러한 연관규칙의 유용성을 측정하는 지표로는 지지도, 신뢰도 및 관심도 등이 있다[1].

전통적으로 연관규칙 탐사에서는 사용자가 정의한 지지도 및 신뢰도 하한값을 만족하는 연관규칙들을 모두 찾아내는데 초점을 맞추어왔으며, 현재는 잘 알려진 apriori 알고리즘 및 그 변종들이 널리 활용되고 있다. 문제는 지지도 및 신뢰도 하한값에 따라 차이는 있으나, 일반적으로 알고리즘 실행 결과 지나치게 많은 연관규칙들이 추출될 수 있어, 이들을 해석하고 활용하는 것이 까다로울 수 있다는 점이다.

이에, 연관규칙 탐사 결과를 보다 활용하기 용이한 형태로 시각화(visualization)하기 위한 연구가 다양하게 수행되어 왔으며[2], 본 논문에서는 구조화된 연관맵이라는 새로운 형태의 연관규칙 시각화 방법을 제안하고자 한다.

Ⅱ. 구조화된 연관맵

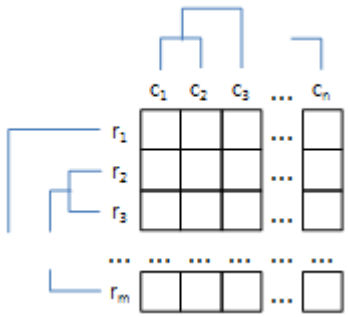
연관규칙 탐사 결과를 요약하는 가장 기본적인 방법은 추출된 연관규칙들을 표에 나열하고, 이들을 지지도, 신뢰도 및 관심도 등의 지표 기준으로 정렬하는 것이다. 보다 발전된 시각화 방법들에서는 평행좌표(parallel coordinate), 네트워크나 차트 및 행렬(matrix) 등이 이용되기도 한다.

이 중 행렬 기반 시각화는 트랜잭션 데이터를 구성하는 항목들을 행렬의 행과 열에 배치하고, 행렬의 각 원소에는 행항목→열항목에 해당하는 연관규칙의 유용성 척도를 대응시키는 것을 의미한다. 가장 대표적으로는 행과 열 항목으로 각각 규칙 전항과 후항의 항목들을 배치하고, 각 원소에 해당 연관규칙의 신뢰도를 대응시킨 다음, 색깔을 이용하여 신뢰도가 높은 경우와 낮은 경우를 구분하는 방법 등이 있다. 이러한 행렬 기반 시각화는 특히 그 내용이 매우 직관적이고 이해하기 쉬우나, 전항과 후항 모두 1개 항목으로만 구성된 연관규칙들에 대해서만 시각화가 가능하다는 단점이 있다. 이를 해결하기 위해서는 행과 열에 등장하는 항목들을 보다 의미 있는 순서로 정렬해야할 필요가 있다.

한편, 행렬 기반 시각화는 연관규칙 탐사 이외에 일반적인 데이터 시각화 목적으로도 널리 활용되어 왔으며, 이러한 경우에도 행과 열 항목 정렬의 필요성이 존재하였다. 다만, 일반적인 데이터 시각화에서는 행렬의 각 원소에 대응되는 값이 인접 원소들과 유사하게 만들거나, 주 대각선(main diagonal) 인근에만 큰 값이 나타나도록 하는 것이 목적이었고, 이에 따라 발견적 기법이나 특이값 분해(singular value decomposition) 등과 같은 선형대수적 분석이 활용되었다[3]. 반면, 연관규칙을 시각화하는 경우에는 단순히 원소의 값 편차를 적게 만드는 것 이외에, 행렬에서 직접적으로 드러나지 않는 다대다 관계의 연관규칙 탐사가 용이해야 한다는 문제가 생긴다. 예를 들어, 두 개의 연관규칙 $A \rightarrow B$, $C \rightarrow D$가 모두 유용하고 신뢰도가 거의 유사하다고 하더라도 항목 A와 C, 또는 B와 D 사이의 상관관계가 높지 않은 경우에는 굳이 전항의 항목 A와 B, 후항의 항목 C와 D를 인접하게 배열할 이유가 없으며, 오히려 이들은 서로 멀리 배치하

는 것이 나을 수도 있다.

한편, 행렬 기반 시각화 방법의 개선을 위해 클러스터 히트맵(cluster heat map)이 사용되는 경우도 있다[4]. 클러스터 히트맵은 그림 1과 같은 구조를 가지며, 행 항목과 열 항목에 항목 간 관계를 나타내는 계통도가 추가된 것이 특징이다. 계통도는 계층형 군집 분석을 통해 얻을 수 있으며, 이를 통해 다양한 항목들 간의 연관성을 도식적으로 표현해줄 뿐만 아니라, 행 항목 및 열 항목을 어느 정도 의미 있게 정렬하는 효과까지 얻을 수 있어, 특히 최근 분자생물학 분야에서 microarray 데이터를 표현하는데 많이 사용되고 있다.



▶▶ 그림 1. 클러스터 히트맵 구조

이러한 클러스터 히트맵을 사용하기 위해서는 먼저, 행 항목-열 항목 간 관련성을 정의해야 한다. 이 부분은 연관규칙 탐사에서 비교적 쉽게 해결할 수 있으며, 일반적으로 행 항목을 전항, 열 항목을 후항으로는 하는 연관규칙의 유용성 지표, 예를 들어 신뢰도나 관심도를 사용하는 경우가 많다. 두 번째로는 행 항목 집합 및 열 항목 집합 내 원소 간 유사도 도는 비유사도가 정의되어야 하는데, 트랜잭션 데이터 내 항목 A, B간 유사도로는 (1)과 같이 항목 집합 지지도에 해당하는 공기 정보(co-occurrence)를 사용하는 경우가 많으며, 비유사도로는 (2)와 같은 자카드(Jaccard) 거리를 사용할 수 있다. 단, $|i|$는 항목 집합 i를 포함하는 트랜잭션의 개수를 의미하며, $|T|$는 트랜잭션에 포함된 모든 레코드 개수를 의미한다.

$$SIM(A,B) = SUPPORT(A \cup B) = \frac{|A \cup B|}{|T|} \quad (1)$$

$$DIST(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

그러나, 연관규칙 탐사에 클러스터 히트맵을 적용할 때, (1)이나 (2)를 일괄적으로 적용하는 것에 대해서는 여전히 의문이 남는다. 규칙 전항을 형성하는 행 항목의 경우, 공기 정보에 기반한 계통도를 통해 빈발 항목 집합을 형성할 가능성이 높은 것들을 인접한 위치에 배치하는 효과가 있지만, 규칙 후항을 형성하는 열 항목의 경우에는 열 항목 내에서의 연관성보다는 행 항목에 있는 규칙 전항과의 연관성이 고려되어야 하기 때문이다.

이에 본 논문에서는 클러스터 히트맵을 연관규칙 탐사 결과 표현에 적합하도록 맞춤화한 구조화된 연관맵을 제안하고자 한다. 구조화된 연관맵의 생성 절차는 다음과 같다.

1) 공기 정보에 기반한 계층형 군집 분석을 통해 행 항목 r_i (i=1, 2, …, m)들에 대한 계통도를 생성한다.
2) 계통도에 기반하여 행 항목을 정렬한다.
3) 열 항목 c_j (j=1, 2, …, n)에 대하여 공기 정보 벡터 cv_j를 작성한다.
4) cv_j에 기반한 계층형 군집 분석을 통해 열 항목들에 대한 계통도를 생성한 후, 이들을 정렬한다.
5) 정렬된 행 항목, 열 항목에 맞게 행렬 내 원소들을 완성한다.

Ⅲ. 결론

구조화된 연관맵의 가장 큰 특징은 열 항목의 구성 방식에 있으며, 공기 정보 벡터 cv_j에 기반하여 계통도를 생성하기 때문에 행 항목들에 대해 비슷한 발생 경향을 갖는 항목들을 인접한 곳에 배치시키는 효과가 발생한다. 따라서, 연관규칙 탐사 결과 추출된 유용한 규칙들의 후항을 보다 체계적으로 관찰할 수 있다. 나아가, 분석자들은 이를 통해 추출된 연관규칙들의 특성을 보다 편리하게 파악할 수 있을 것으로 기대된다.

▌감사의 글▐

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임 (2012R1A1A1044834)

▌참 고 문 헌▐

[1] Tan, P.-N., Steinbach, M. and Kumar, V., Introduction to Data Mining, Addison-Wesley, 2005.
[2] Fernandes, L.A. and Garcia, A.C.B, “Association rule visualization and pruning through response-style data organization and clustering” In Advances in Artificial Intelligence-IBERAMIA, pp.71-80, 2012
[3] Liiv, I. “Seriation and matrix reordering methods: An historical overview”, Statistical Analysis and Data Mining: The ASA Data Science Journal, Vol.3, No.2, pp.70-91,2010.
[4] Wilkinson, L. and Friendly M. “The history of the cluster heat map”, The American Statistician, Vol.63, No.2, pp.179-184, 2009.
[5] Day, W.H.E. and Edelsbrunner, H. “Efficient algorithms for agglomerative hierarchical clustering method”, Journal of Classification, Vol.1, No.1, pp.7-24, 1984.