데이터 마이닝을 이용한 의사결정지원 시스템

조성진, 정인정
고려대학교 전산학과

Decision Support System Using Data Mining

Sung-Jin Cho, In-Jeong Chung
Dept. of Computer Science, Korea Univ.

요 약

데이터 베이스에 저장하고 취급하는 자료가 폭발적으로 증가함에 따라서, 데이터 베이스 이용자가 필요로 하는 자료를 검색하고 유용한 정보를 획득하는 일은 더욱 더 어려워지고 있다. 이러한 문제들은 데이터에 내재되어 있는 유용한 패턴이나 변수들 간의 관계를 정교한 분석 모형을 찾아내는 데이터 마이닝이란 정보기술로 해결할 수 있다. 본 논문에서는 여러 가지 데이터 마이닝 기법들을 알아보고 데이터 마이닝에 의해 만들어진 규칙들을 사용하여 의사결정에 도움을 줄 수 있는 분석적인 트리를 구성한다. 제안하는 트리가 어떻게 생성되는지 보이고 생성된 트리를 의사결정지원 시스템에 적용한다. 다양한 관점에서 분석을 요구하는 사용자를 충족 시키는 트리를 구성하여 시각적인 효과와 각 계층간의 분석을 할 수 있는 의사결정지원 시스템을 소개한다.

1. 서론

기존의 데이터 베이스 관리 시스템이 취급하고 이에 저장되는 데이터의 양이 증가함에 따라서 저장된 자료로부터 원하는 정보를 효과적으로 추출하거나 데이터에 내재된 규칙들을 찾는 일이 대단히 어렵게 되고있다.[7] 우리는 1980년 중반에 소개된 데이터 마이닝 이라는 유용한 정보기술을 이용하여 기존의 이와 같은 데이터 베이스 관리 시스템의 문제점들을 해결할 수 있을 뿐만 아니라 내재된 정보 및 규칙에서 새로운 지식을 얻을 수 있다.

본 논문에서는 데이터 마이닝의 유용한 정보기술을 이용하여 많은 잠재되어있는 유용한 정보들을 마이닝 함으로써 의사결정지원 시스템의 효율성을 높이고자 한다.

본 논문의 구조는 다음과 같다. 2장에서는 데이터 마이닝의 개념부터 시작하여 데이터 마이닝의 여러 가지 기법, 그리고 관련기술을 고찰하고 3장에서는 의사결정지원 시스템의 개념 및 구조 등을 살펴본다. 4장에서는 데이터 마이닝 기법을 이용한 의사결정 지원시스템을 제안하고 5장에서는 실례를 들어 타당성을 검증하고 마지막으로 6장에서는 결론을 살펴본다.

2. 데이터 마이닝

데이터 마이닝을 효과적으로 수행하기 위해서는 많은 사전 및 사후 작업이 필요하다. 즉, 어떠한 데이터가 마이닝될 필요가 있는지 적절한 데이터를 준비하고, 마이닝에 적합한 형태로 데이터를 가공하고, 마이닝에 사용할 기법을 선택하고, 마이닝된 결과를 해석하고, 의사결정에 활용하는 일련의 과정을 필요로 한다 [7]. 이런 전반적인 과정을 KDD(Knowledge Discovery in Database)라고 한다.

데이터 마이닝 (Data Mining)은 데이터 베이스에 저장된 데이터에서 유용한 정보를 추출하는 작업이다. 또한 데이터 마이닝은 데이터에 내재되어 있는 유용한 패턴이나 변수들간의 관계를 정교한 분석 모형을 사용하여 찾아내는 작업으로 기업들이 보유한 기존의 경험적 지식을 재확인하는 역할을 수행함과 동시에 지금까지 인식하지 못했던 새로운 정보를 제공하여 경영의사결정에 도움을 주는 정보기술이다.

이러한 데이터 마이닝 기법에는 의미 있는 몇 개의 군집으로 나누는 클러스터링 (Clustering) 기법 [7], 다른 클래스에 대한 차별적 특성을 도출하는 분류 (Classification) 기법 [7,11], 관련성을 구하는 연관 규칙 (Association Rule) 기법 [7,11], 일반화된 대표적 표현으로 축약시키는 요약 (Summarization) 기법 [7]등 다양한 기법들이 있다.

3. 의사결정지원시스템

3.1 의사결정지원 시스템의 개념

1970년대 초기에 Gorry와 Scott-Morton에 의해서 경영의사결정시스템 (Management Decision System)이라는 용어로 정립되기 시작하였다. "의사결정자가 비정형적 문제를 해결하기 위해서 데이터와 모형을 사용하도록 도와주는 상호작용적인 컴퓨터기반 시스템"을 의사결정지원 시스템 (Decision Support System : DSS)으로 정의하였다[2, 3].

3.2 의사결정지원 시스템의 구조

의사결정지원 시스템은 여러 개의 하부시스템이 유기적으로 결합된 구조를 갖는데, Sprague와 Carlson에 의해서 제시된 데이터관리, 모형관리 및 대화관리의 3개의 하부시스템으로 구성된 전통적인 의사결정지원 시스템의 구조가 있다[4].

1980년대 중반에 인공지능 분야의 발전은 의사결정지원 시스템에도 많은 영향을 주었다. 특히 전문가 시스템 (Expert System) 또는 지식기반 시스템 (Knowledge-Based System)과의 결합은 의사결정지원 시스템에 전문지식 (Expertise Knowledge)과 추론기능을 추가하였다[5, 6]. 이 시스템은 전통적 구조와 구분하기 위해서 지능형 의사결정지원 시스템 (Intelligent DSS) 또는 지식기반 의사결정지원 시스템 (KB-DSS)이라는 용어를 사용한다.

4. 데이터 마이닝을 이용한 의사결정지원 시스템

4.1 데이터 마이닝과 의사결정지원 시스템의 관계

데이터 마이닝을 통해 유용한 정보를 추출한다는 자체가 분석적인 데이터를 얻기가 용이하다는 것이다.

- 데이터 마이닝은 의사결정지원 시스템의 기능을 확장할 수 있는 새로운 정보기술로 간주할 수 있다.
- 데이터 마이닝이 지닌 지식 집약적이고 상호작용적인 특성과 최종사용자의 지원이라는 목표는 의사결정지원 시스템과 동질적이기 때문에 의사결정지원 시스템 프레임워크를 적용하여 데이터 마이닝 통합환경 (Integrated Environment for Data Mining : IEDM)을 구현할 수 있다.

4.2 특성화 규칙 (Characteristic Rule)을 이용한 의사결정지원 시스템

4.2.1 의사결정 트리 (Decision Tree)

하나의 주어진 상황속에서 취할 수 있는 대처 방안을 제시하고, 각각의 조건이 발생한 경우에 진행할 수 있는 행동들을 트리 형식으로 표현한 것이다 [13].

의사결정 트리를 구하는 대표적인 알고리즘으로 ID3 (Interactive Dichotomizer 3)가 있다. ID3는 Quinlan이 개발한 학습방법으로서 어떤 개념에 관한 예와 반례로써 훈련집합이 주어졌을 때 이로부터 개념을 구별할 수 있는 의사결정 트리 형태의 분류규칙을 생성시킨다. 여기서 분류하고자 하는 개념들을 클래스(Class)라 하고, 이 클래스에 관한 예는 해당 클래스를 한정된 수의 특성으로써 묘사된다 [12].

4.2.2 특성화 규칙 (Characteristic Rule)

특성화 규칙은 각 객체들의 집합을 일반화시키고 요약, 간결하게 표현함으로써 상대적으로 상위 레벨의 규칙들을 표현하는 방법이다 [1]. 먼저 상위 튜플들을 전체에 대해 비율을 구하고 그 다음 레벨의 튜플들을 상위 튜플에 대해 비율을 구한다.

특성화 규칙을 생성하는 단계는 원시 테이블과 같은 원시 클래스를 표현하고 일반화 된 Crosstab 형태의 원시 클래스를 표현하고 원시 클래스에 있는 속성들을 축소시킨다.

지역	상품	연령	소득수준	판매량
A	1	20대	하	10
A	1	20대	중	15
...	...	...	...	...
C	3	50대 이후	상	8

표 1 원시 테이블

먼저 다음 표 1과 같은 원시 테이블이 주어지면 지역, 상품, 연령, 소득수준의 4개의 차원으로 표현된 원시 테이블을 표 2와 같은 Crosstab 테이블로 다시 일반화 시킨다.

Crosstab 테이블에서는 원시테이블의 각각의 튜플들로 t-weight라 불리는 논리적인 규칙들을 mapping시킬수 있다.

지역	상품	20대				30대				40대				50대 이후				계
		하	중	상	계	하	중	상	계	하	중	상	계	하	중	상	계	
A	1	10	20	40	70	18	15	28	61	12	22	16	50	13	15	22	50	231
	2	15	35	20	70	8	22	20	50	5	18	5	28	12	10	8	30	178
	3	20	8	10	38	11	24	15	50	18	10	8	36	16	8	10	34	158
	계	45	63	70	178	37	61	63	161	35	50	29	114	41	33	40	114	567
B	1	5	15	38	58	19	10	29	58	7	23	18	48	10	13	30	53	217
	2	13	42	22	77	10	18	18	46	9	15	12	36	9	12	9	30	189
	3	21	12	13	46	9	20	11	40	13	14	11	38	15	15	5	35	159
	계	39	69	73	181	38	48	58	144	29	52	41	122	34	40	44	118	565
C	1	12	10	29	51	11	10	30	51	8	15	16	39	8	18	28	54	195
	2	8	29	8	45	9	20	15	44	8	10	10	28	7	12	6	25	142
	3	22	8	10	40	7	21	9	37	15	13	8	36	11	11	8	30	143
	계	42	47	47	136	27	51	54	132	31	38	34	103	26	41	42	109	480
계		126	179	190	495	102	160	175	437	95	140	104	339	101	114	126	341	1612

표 2 지역별, 연령별 상품 판매량을 나타낸 예제 테이블

q_a를 일반화 된 튜플들이라 하고 q_a에 대한 t_weight는 전체에서 q_a가 차지하는 비율로 나타내면 다음과 같다.

$$t\_weight = \frac{count(q_a)}{\sum_{i=1}^{n} count(q_i)}$$

n은 튜플들의 개수이고 일반화된 q_a 는 $q_1,...q_n$.

위에서 살펴본 t-weight를 가지고 이제 규칙들을 생성하면 다음과 같다. 표 2에서 볼 수 있듯이 각각의 튜플들은 다차원적으로 구성되어 있고 또한 서로가 연관이 있다. 예를들면 지역만 살펴보면 A지역 B지역 C지역으로 나누어져 있고 이들 지역간의 판매량이 나타난다. 하나의 튜플만으로 규칙을 생성하면 A지역의 판매량은 간단히 나온다. 전체 판매량에서 A지역이 차지하는 비율을 구하면 바로 나올수 있는 것이다. 즉, 지역_A = 35.17% 라는 규칙을 생성할 수 있다.

이제는 지역과 상품의 두 개의 튜플을 동시에 살펴보면 두 개의 튜플은 연관성을 가지고 있다. 지역 A에서의 상품 1의 판매량으로 규칙을 생성하면, 지역_A ^ 상품_1 = 11.04%가 된다. 생성된 규칙은 A지역에서 상품 1의 판매량을 전체 판매량에 대해 분석한 것이다. 다시 살펴보면 A지역에서 35.18%의 판매량이 있는데 그 중에 상품 1은 11.08%의 비율을 차지하고 있는 것이다.

다른 각도에서 상품1의 판매량을 먼저 분석하고 다음 지역을 분석하면 다음과 같은 규칙이 생성된다.

상품_1 = 39.89%, 상품_1 ^ 지역_A = 14.43%. 생성된 규칙을 살펴보면 상품 1이 전체 판매량에 39.18%를 차지하고 있고 그 중에 A 지역에서 상품 1의 판매량은 14.43%가 되는 것이다.

이러한 절차를 통해 먼저 분석하고자 하는 튜플을 정하고 그 튜플의 특성화 규칙을 생성하고 그 다음 단계의 튜플들을 정하고 또 특성화 규칙을 생성하면 된다. 중간에 생성된 규칙들은 그 규칙들간의 분석적 데이터로써 의미가 있다. 이렇게 생성된 규칙들은 제안하는 의사결정 지원시스템의 트리를 구성하는 데이터로 활용하게 된다.

4.2.3 제안하는 의사결정지원 시스템

우리는 본 논문에서 특성화 규칙을 트리구조로 나타내어 기존의 의사결정 트리보다 더 효율적으로 의사결정을 구현하고자 한다. 일반적인 의사결정 트리와 다르게 본 논문에서 제안하는 트리는 분석적인 데이터를 보다 효과적으로 한눈에 들어올 수 있는 형상화된 도표로서 의사결정을 하는데 도움이 되는 자료를 만들고자 하는 것이다.

특성화 규칙을 생성한 후 다음과 같은 절차에 의해서 트리를 구성한다.

1. 일단 가장 상위에서 분석하고자 하는 튜플을 선정한다.
2. 선정된 튜플을 전체 판매량의 자식노드로 놓고 생성된 특성화 규칙을 표기한다.
3. 다음 단계에 분석하고자 하는 튜플을 선정한다.
4. 1-3번까지 반복하고 더 이상 분석할 튜플이 없을 경우는 트리가 완성되는 것이다.

상위 레벨의 튜플이 일정하게 주어진 것이 아니라 각각의 튜플들을 최상위 레벨에 놓고 분석해 갈 수 있도록 한다. 각각의 튜플들의 특성화 규칙에 의해서 어떠한 튜플이라도 상위 레벨에서 분석되어질 수 있다. 다시 말해서 여러 방면에서 분석이 가능하다.

기존의 ID3[12,13]는 각 단계에서 생성된 트리들은 버리고 마지막 트리만을 생성하는데 비하여, 본 논문에서 제안하는 트리는 어떠한 튜플들을 먼저 선택하고 트리를 구성하더라도 생성된 트리는 모두 분석적인 도구로써 활용할 수 있다. 또한 제안하는 트리는 여러 각도에서 의사결정용 분석트리를 구성함으로써 의사결정을 하는데 많은 도움을 줄 수 있다.

5. 타당성 검증

4장의 예제 테이블인 표2는 각각의 테이블의 튜플들간에 관계를 다차원 관계로 표현한 테이블들이다. 각 튜플들간의 숫자들은 판매량을 나타내고 있고 판매량을 전체 판매량으로부터 각 판매량의 분석을 통해 트리구조로 나타내어 의사결정에 도움이 될 수 있는 도표를 나타내고자 한다. 먼저 이들 각각의 튜플들중 지역을 최상위 레벨로 하여 구한 특성화 규칙은 다음과 같다.

지역_A = 35.17%
지역_B = 35.05%
지역_C = 29.78%
지역_A ^ 상품_1 = 11.04%
지역_A ^ 상품_1 ^ 연령_20대 = 4.34%
지역_A ^ 상품_1 ^ 연령_20대 ^ 소득수준_하 = 0.62%
. . .

위의 특성화 규칙을 이용하여 그림 1과 같은 트리를 만들어 낼 수 있다. 지금의 트리에서는 지역을 최상위 레벨로 하여 분석한 트리로 예를 들면 어떤 기업에서 여러 지역에 지사가 있다고 가정할 때 지역별로 판매량을 분석할 때 용이하도록 분석한 트리이다. 지역을 최상위 레벨로 하여 분석할 수도 있지만 소비자별로 또는 상품별로 트리를 재구성할 수도 있다.

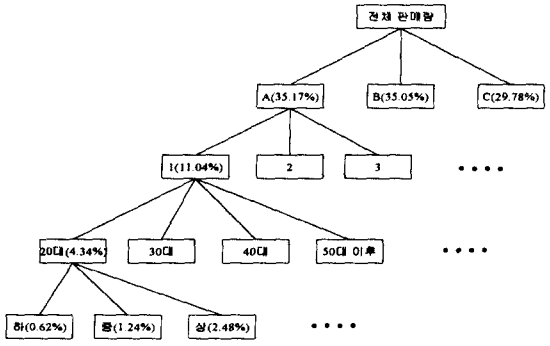

그림 1 특성화 규칙을 이용하여 구성한 트리

그림 1을 보고 트리를 분석해보면 우선 전체 판매량에 대해 지역별로 어떤 지역이 가장 실적이 좋은지를 구분 할 수 있게 된다. 그 다음 지역안에서도 어떤 제품이 가장 선호도가 좋았는지, 선호도가 좋은 제품중에서 어느 연령층이, 소득수준은 어떤 상태인 고객들이 그 제품을 많이 구매했는지를 단계적으로 한눈에 알아 볼 수 있다. 즉, A지역이 전체 판매량에 대해 35.17%의 실적을 보이고 있는데 A지역에서 판매한 제품 중 상품 1이 20대의 소득수준이 상인 고객들로부터 가장 선호도가 좋았다는 것을 알 수 있다. 이러한 것들은 판매 전략을 세우는데 효율적인 정보로 사용될 수 있다. 지역별로 분석한 후 어떤 지역은 어떤 상품이 선호도가 좋다라는 평가가 나오면 선호도가 좋은 상품은 더 많은 생산을 함으로써 효과적인 기업 운영을 할 수 있다.

기존의 의사결정 트리와는 달리, 본 논문에서 제한하는 방법은 분석적인 데이터를 트리구조로 나타내어 필요한 정보로 한눈에 모든 상황을 살필 수 있고 트리구조로 되어있기 때문에 계층적으로 요소별로 분석이 가능하다. 또한 고정된 관점에서의 분석만을 제공하는 것이 아니라 같은 테이블이 주어졌을 때 여러각도에서 분석할 수 있는 다양한 트리를 제공한다는 점 또한 장점이다.

6. 결론

데이터 마이닝은 데이터 베이스에 저장된 대량의 데이터로부터 필요한 정보를 추출하고 내재된 규칙들은 도출하는 정보기술이다.

본 논문에서는 의미 있고 의사결정에 도움이 될 수 있는 지식을 획득하는 데이터 마이닝 기법을 사용하여 의사결정지원 시스템을 구성하였다. 데이터 마이닝 기법중에서도 특성화 규칙을 사용하여 규칙을 생성하고 분석 트리를 구성하였고 계층적으로 분석이 용이한 트리구조를 선택하였다. 제안하는 트리는 기존의 의사결정 트리와 다르게 다각도로 분석한 결과를 시각화 시킴으로써 사용자가 원하는 만큼 세분화하여 각 요소별로 다양한 분석을 할 수 있다. 특히 같은 테이블로 여러 각도에서 분석할 수 있는 트리를 제공함으로써 다양한 사용자 관점에서 분석할 수 있다는 것이 장점이다.

참고 문헌

[1] Jiawei Han, Shojiro Nishio, Hiroyuki Kawano, and Wei Wang, "Generalization-based data mining in object-oriented databases an object cube model", Data & Knowledge Engineering, Vol 25, pp. 55-97, 1998.

[2] Gorry, G. A., and M. S. Scott-Morton, "A Framework for Management Information Systems," Sloan Management Review, Vol. 13, No. 1, pp. 55-70, Fall 1971.

[3] Scott-Morton, M. S., Management Decision Systems : Computer Based Support for Decision Making, Cambridge, MA : Division of Research, Harvard University, 1971.

[4] Sprague, R. H., Jr., and E.D. Carlson, Building Effective Decision Support Systems, Englewood Cliffs, NJ : Prentice-Hall, 1982.

[5] Klein, M., and L. B. Methlie, Expert Systems : A Decision Support Approach, Wokingham : England, Addison-Wesley, 1990.

[6] Turban, E., and P. R. Watkins, "Integrating Expert Systems and Decision Support Systems", MIS Quarterly, Vol, 10, No. 2, pp. 121-136, 1986.

[7] M.-S. Chen, J. Han and P. Yu, "Data Minning : An Overview from Database Perspective", IEEE Trans. on Knowledge and Data Engineering, 1997.

[8] W.H. Inmon, "Building the Data Warehouse", Second Edition, John Wiley and Sons. Inc., 1996

[9] Codd, E.F., S.B. Codd, and C.T. Salley, "Providing OLAP to User-Analysts : An IT Mandate", White Paper, Codd & Date Inc. 1993.

[10] Dorrian, Jim, OLAP- The Multi-dimensional Approach to Data Analysis, Mini-Micro Systems, April 1994, pp. 10-11.

[11] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, "Database Mining : A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 5, No. 6, December 1993, pp. 914-925.

[12] Patrick Henry Winston, "Artificial intelligence", 3rd ed., Patrick H. Winston., 1992.

[13] Cezary Z. Janikow, "Fuzzy Decision Trees : Issues and Methods", IEEE Transactions on Systems, Man, and Cybernetics-Part B : Cybernetics, Vol. 28, No. 1, Feb., 1998.