

# 시계열 예측에 대한 의사결정자의 인지 유형과 생리적 반응 특성의 상관분석을 위한 데이터 마이닝 접근방법

송병호<sup>+</sup>, 박흥국<sup>\*\*</sup>  
<sup>+</sup> 상명대학교 소프트웨어학과  
<sup>\*\*</sup> 상명대학교 정보통신학부

## Data Mining Approach to Analyzing the Effect of Cognitive Style and Physiological Phenomena in Judgemental Time Series Forecasting

Byoungho Song<sup>+</sup>, Hung Kook Park<sup>\*\*</sup>  
<sup>+</sup> Dept. of Software Science, Sangmyung University  
<sup>\*\*</sup> Division of Information and Telecommunications, Sangmyung Univ.

데이터 마이닝이란 축적된 방대한 양의 실제 데이터로부터 이전에는 알지 못했던, 숨겨진 임의의 규칙성들을 비 전통적인 방식으로 발견해 내는 작업을 말한다. 많은 데이터로부터 무엇인가 흥미로운 경향이나 패턴을 발굴해 내는 것이 데이터 마이닝의 목적이다. 본 연구에서는 다양한 측정값으로 표현되는 ① 인지 유형 데이터와, ② 생리적 반응 특성 데이터가 ③ 직관적 예측의 성과에 미치는 영향을 데이터 마이닝 기술을 이용하여 분석함으로써 존재하는 규칙적인 관련성을 탐사하였다.

현재까지 분석한 바로는 첫째, 분석적인 사람이 직관적인 사람보다 예측이 더 정확한 경향이 있었다. 둘째, 실험 전과 실험중 간의 뇌파증가율이 높거나 뇌파량이 적으면 분석적인 사람일 가능성이 많았다. 셋째, 분석적인 사람은 실험 전에 뇌파량이 적을수록 더 정확해지며, 직관적인 사람은 실험 전에 뇌파량이 많을수록 더 정확해지는 것으로 관측되었다.

### 1. 서론<sup>1)</sup>

시계열 직관 예측 (judgemental time series forecasting)은 일상 생활에서 쉽게 대하는 문제이다. 주가 분석이나 일기 예보 등이 그 예가 된다. 본

연구팀에서는 의사결정자의 인지 유형과 생리적 반응 특성이 직관적 예측의 성과에 미치는 영향을 분석하기 위하여 여러가지 실험을 하였다. 그런데 그 측정치들을 눈으로 보아서는 상관관계가 쉽게 나타나지 않기 때문에 많은 실제 데이터로부터 눈에 보이지 않는 흥미로운 특성들을 탐사해 내는 최신 기술인 데이터 마이닝을 적용해 보았다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 데이터 마이닝에 대하여 간단히 소개한다. 3장에서

1) 이 논문은 1999년도 한국학술진흥재단의 연구비에 의하여 연구되었음. (KRF-99-번호미지정. 제목: 시계열 직관 예측에 영향을 주는 의사결정자의 인지적/생리적 특성분석에 관한 연구)

는 의사 결정자들의 인지 유형, (뇌파로 측정되는) 생리 특성과 실제 시계열 예측시의 오류율을 30명을 대상으로 실험한 내용에 대하여 설명한다. 4장에서는 이러한 데이터를 바탕으로 이 연구에서 추진중인 데이터 마이닝 방법을 기술한다. 현재까지 진행된 데이터 마이닝의 결과를 5장에서 분석하며, 마지막으로 6장에서 결론을 맺는다.

## 2. 데이터 마이닝 소개

데이터 마이닝(data mining: 데이터 발굴, 데이터 탐사)이란 축적된 방대한 양의 실제 데이터로부터 이전에는 알지 못했던, 숨겨진 임의의 규칙성들을 비 전통적인 방식으로 발견해 내는 작업을 말한다. 컴퓨터를 이용한 정보 활동의 급격한 증가와 발전에 따라 인간이나 전통적인 시스템이 이해하고 활용하기에는 너무 많은 데이터들이 모이게 되었고, 이렇게 모인 데이터로부터 무엇인가 흥미로운 경향이나 패턴을 발굴해 내는 것이 데이터 마이닝의 목적이다.

예를 들어, 신용 카드 사용 내역을 분석해서 분석되었거나 도난당한 카드를 사용하는 것이라는 사실을 탐지하는 응용, 고객에 관한 기록들을 뒤져서 신상품에 흥미를 보일 만한 고객들을 골라 내어 홍보에 이용하고 싶은 응용, 마찬가지로 고객의 연령, 소득 수준, 결제 패턴 등을 분석하여 고객 신용도를 예측하는 응용 등에서 이 데이터 마이닝 기술은 괄목할 만한 성과를 보여주고 있다[1-7].

마이닝의 가장 기본적인 예가 장바구니(market basket) 분석이다. 고객이 한 번 쇼핑할 때 같이 구매하는 경향이 있는 품목들을 찾는 것이 대표적인 탐사 목적이다. 예컨대 펜을 사는 사람이 잉크도 같이 사는 경향이 많다고 하면,

펜 → 잉크

라는 규칙(rule)이 발견된다. 그런데 고객들은 필기도구를 한 번에 모아서 사는 경향이 많다고 하면,

펜 → 연필

이라는 규칙도 발견될 것이다. 이 두 가지 규칙에 따라

연필 → 잉크

도 상당한 수준으로 관측될 수 있다. 그러나 이 규칙은 우리가 원하는 근본적인 상관관계가 아니다. 이 잘못된 관찰 결과에 바탕하여, “연필을 사면 잉크를 할인해 준다”는 판매 전략을 세우더라도 별 소득이 없게 된다. 그러므로 데이터 마이닝을 할 때에는 가장 근본적인 인과관계를 찾는 것이 중요하다. 이 일은 쉽지 않은 작업이다.

발견된 규칙을 평가하는 기준으로는 지지도

(support)와 신뢰도(confidence)가 있다. 의미가 있는 규칙이라면 우연히 한두 번 나타나는 것이 아닐 것이다. 이 연구는 마이닝 중에서도 분류 규칙(classification rule)을 찾는 마이닝에 해당하는데, 이때에는 규칙의 좌측 항이 모두 나타나는 사실이 전체 사실중에서 차지하는 백분율을 지지도라고 한다. 그러한 사실들 중에서 우측 항의 조건을 만족하는 백분율을 신뢰도라고 한다.

데이터 마이닝의 절차는 일반적으로 다음과 같이 여섯 단계로 이루어진다.

1단계 데이터 선정: 모아 놓은 데이터 집합으로부터 분석 대상 데이터들을 선별한다.

2단계 데이터 세정(cleaning): 잘못된 부분을 바로잡는다. 예를 들어 동일한 사실을 여러 번 입력하였거나 같은 유형의 데이터를 다른 표기법으로 표시한 것들을 바로잡는다.

3단계 보강: 필요한 경우에는 새로운 데이터들을 추가하여 마이닝 효과를 최대화한다.

4단계 부호화: 최적의 구조로 데이터를 조직한다. 보강된 데이터를 융합시키고, 데이터 값들을 상/중/하 등 구간별로 분류하여 탐사를 준비한다.

5단계 실제 탐사 작업: 여러가지 도구와 방법을 이용하여, 전 단계에서 최종적으로 준비한 데이터 집단에 대하여 마이닝 작업을 실시한다.

6단계 보고서 작성: 탐사된 결과를 분석하여 최종 결과를 도출한다.

## 3. 연구 환경

이 연구는 학술진흥재단의 협동연구과제 “시계열 직관 예측에 영향을 주는 의사결정자의 인지적/생리적 특성 분석에 관한 연구”의 예비 연구로 진행되었다. 모 연구에서는 의사결정자에게 임의적으로 추출된, 주가 동향이나 시장 조사와 같은 표준 시계열 추이 패턴을 주고, 그 이후가 어떻게 될 것인지를 판단하도록 하는 실험을 지난 5월 30명의 학생을 대상으로 실시하였다. 이때 각 피험자의 인지 유형을 몇 가지 시험을 통하여 T(분석적)와 F(직관적, 감성적)의 두 가지로 분류하였는데, 각 15명씩이었다. 또 각 피험자의 생리적 특성을 구하기 위하여 실험하기 전과 실험하는 도중 좌뇌와 우뇌에서 발생하는  $\alpha$  파와  $\beta$  파를 각각 측정하였다. 이렇게 해서 구한 실험 데이터는 표 1과 같다. 프라이버시를 위하여 여기에서는 피험자들의 신원을 감추었다. MAPE는 예측 오류율을 나타낸다.

이 데이터 자체만을 놓고 볼 때 별다른 유의성을 발견할 수 없었다. 따라서 데이터 마이닝 기법을 이용하여 얼른 눈에 띄이지 않는 규칙들을 찾아 보기



유도할 수 있는 값은 이 밖에도 많기 때문에 이들을 효율적으로 계산해 내고 효과적으로 조직할 수 있는 데이터베이스 설계가 필요하다. 이 논문에서는 우선 주어진 실험 데이터의 항목을 기준으로 마이닝을 추진해 보고자 하였다.

다음으로, 세정 과정과 보강 단계는 필요 없다.

각 데이터 항목에 대한 값 범위의 구간 분류는 최종적인 목표가 각각을 3-5개 구간으로 나누어 보는 것이지만, 현재로서는 실험 수가 적어서 이렇게 되면 유의성 있는 규칙을 발견할 수가 없었다. 따라서 이 논문에서는 각각을 고/저의 두 구간으로 나누어 보았다. 또, 구간 구분을 폭에 따라 균등 분할하는 방식과, 도수에 따라 분할하는 방식을 모두 수행해 보는 것이 장기 계획이지만 여기에서는 우선 도수에 따라 균등 분할하기로 하였다. 따라서 실험전 좌뇌  $\alpha$  좌측치는 고 15건, 저 15건이 되는 식이다.

실제 마이닝에는 여러가지 기법이 제안된 바 있다. 장기적으로는 이 연구에 최적인 마이닝 도구를 개발하는 것이 목표이지만, 여기에서는 노르웨이의 Norwegian University of Science and Technology 대학 Dept. of Computer and Information Science 학과에서 1999년 개발한 MS-Windows용 데이터 마이닝 도구 Rosetta를 이용하고, 수작업을 병행하였다. 그런데 데이터 마이닝 기술로 이와 같은 규칙성을 발견해 내려면, 만들어 낼 수 있는 모든 항목 조합별로 관련도를 일일이 계산해 보아서 지지도와 신뢰도가 일정 수준 이상이 되는 것들을 골라 내어야 한다. 이러한 작업은 현대 컴퓨터의 수행 속도가 빠르지만 여전히 많은 시간을 소모하기 때문에, 이 시간을 줄이는 방법을 장래에는 연구할 필요가 있다.

## 5. 마이닝 분석 결과

### 5.1 인지 유형 -> 오류율 관계

인지 유형에 따른 오류율의 차이를 분석해 보았더니, 분석적인 사람(T)의 0.6(9/15)이 오류율이 낮았고 0.4(6/15)가 오류율이 높았다. 반면 직관적인 사람(F)의 0.4(6/15)가 오류율이 낮았으며 0.6(9/15)이 오류율이 높았다.

한편, 오류율이 낮은 사람의 0.6(9/15)이 T형이었으며 0.4(6/15)가 F형이었다. 반면 오류율이 높은 사람의 0.4(9/15)가 T형이었으며 0.6(6/15)이 F형이었다.

### 5.2 뇌파(생리 특성) -> 인지 유형 관계

다음과 같은 형태의 마이닝 규칙에서

좌측항 -> 우측항

좌측항의 길이가 1인 규칙을 단항 규칙, 2 이상인 규칙을 다항 규칙이라고 부르기로 한다. 현재 좌측항에 나올 수 있는 항목의 수를 계산해 보자. 실험 데이터 20개 항목에서 우측항에 나올 인지 유형(style)을 빼고 나면 19개이므로 단항 규칙에서 좌측항에 나올 항목 조합(한 항목으로 구성)의 수는  $19C1=19$ 개이다. 이 각 조합별로 해당 항목이 고/저의 두 값을 가질 수 있으므로, 결국 나올 수 있는 단항 규칙의 수는

$$19C1 * 21$$

이다. 좌측항이 이항, 3항, ... 19항까지 있을 수 있으므로 나올 수 있는 전체 규칙의 수는

$$\begin{aligned} &19C1 * 21 \\ &+ 19C2 * 22 \\ &+ 19C3 * 23 \\ &+ 19C4 * 24 \\ &+ \dots \\ &= 19 * 2 \\ &+ 171 * 4 \\ &+ 969 * 8 \\ &+ 3876 * 16 \\ &+ \dots \end{aligned}$$

등으로 막대한 수가 된다. 그렇기 때문에 지금까지는 좌측항의 길이를 4로 제한해 보았으며, 지지도가 10%(3건/30) 이하인 것들은 제거하였다.

우선 단항 규칙들을 분석한 결과 피험자가 T형이 되는 우측항의 신뢰도는 표 2와 같다.

	$\alpha$			$\beta$		
	좌	우		좌	우	
실험전	↓ 9/15	-		↓ 9/15	-	
실험중	-	-		↑ 9/15	↑ 10/15	
증감율	↑ 10/15	↑ 11/15		↑ 9/15	↑ 10/15	

표 2. 뇌파 -> 인지유형의 단항 규칙 분석표

단항 규칙을 관찰해 보면 다음과 같은 사실을 알 수 있다.

- 실험전 좌뇌의  $\alpha/\beta$  량과  $\alpha/\beta$  좌우합계가 낮으면 T형일 확률이 60%이다.

- 실험중  $\beta$  량이 좌뇌나 우뇌에서 많으면 T형일 확률이 높다(각기 9/15, 10/15). 그러나 좌우합계로는 불분명하다.

- 실험전후 뇌파의 증가량이  $\alpha/\beta$ , 좌/우 막론하

고 많으면 T형일 가능성이 높다.

다항 규칙중 지금까지 관찰된 유의사실은 다음과 같다.

- 좌뇌  $\alpha$  전후증감, 우뇌  $\alpha$  전후증감,  $\alpha$  전후증감좌우합계가 모두 낮으면 F형일 확률이 0.72(8/11)이며, 모두 높으면 T형일 가능성이 0.82(9/11)이다.

-  $\alpha$  전후증감좌우합계와  $\beta$  전후증감좌우합계가 모두 낮으면 F형일 확률이 8/11이며, 모두 높으면 T형일 가능성이 8/11이다.

이로부터 지금까지 내린 잠정 결론은 다음과 같다.

① 뇌파 증가율이 높으면 T형일 확률이 많으며, 낮으면 F형일 확률이 많다.

② 뇌파량이 적으면 T형일 확률이 있다(60% 정도).

### 5.3 뇌파 -> 오류율(정확도) 관계

이 관계 역시 나올 수 있는 규칙의 수는 막대하기 때문에 지금까지는 좌측항의 길이를 4로 제한해 보았으며, 지지도가 10%(3/30) 이하인 것들은 제거하였다.

#### 5.3.1 단항 규칙 분석

우선 단항 규칙들을 분석한 결과 피험자의 오류율이 낮아지는(정확해지는) 우측항의 신뢰도는 표 3과 같다.

	$\alpha$			$\beta$		
	좌	우	합계	좌	우	합계
실험전	-	↑9/15	-	↓9/15	↑9/15	-
실험중	-	-	-	↓9/15	-	-
증감율	-	-	-	-	-	↑9/15

표 3. 뇌파 -> 오류율의 단항 규칙 분석표

단항 규칙을 관찰해 보면 다음과 같은 사실을 알 수 있다.

- 유의성이 잘 나타나지 않는다.

- 비교적 정확도가 높다고 분석된 T형의 생리 특성 경향(표 2)과 일치하는 경향은 실험전좌뇌  $\beta$  파(낮을때)에 불과하다.

- 좌뇌  $\beta$  파가 실험 전이나 실험중에 모두 낮으면 정확도가 높아지는 확률이 있다(60%정도).

- 실험전 우뇌의  $\alpha/\beta$  파가 모두 높으면 정확도가 높아지는 확률이 있다(60%정도).

이렇게 단순한 상관 관계는 찾을 수가 없었다. 따라서 다항 규칙들을 더 자세히 살펴 보아야 한다. 편의상 F형과 T형으로 나누어 분석하였다.

### 5.3.2 F형이 오류율이 낮은 경우

몇 가지 유의성 있는 규칙들을 나열하면 다음과 같다.

- F AND 실험전좌뇌  $\alpha$ (고) 실험전우뇌  $\beta$ (고): 확률 4/5(0.8)

- F AND 실험전우뇌  $\beta$ (고) AND 실험전  $\beta$  좌우합계(고): 5/7(0.71)

- F AND 실험전  $\beta$  좌우합계(고) AND 좌뇌  $\beta$  좌우증감(저): 4/6(0.67)

- F AND 실험전우뇌  $\beta$ (고) AND 우뇌  $\alpha$  전후증감(저): 4/6

- F AND 실험전  $\beta$  좌우합계(고) AND 우뇌  $\beta$  전후증감(저): 4/6

- F AND 실험전우뇌  $\alpha$ (고) AND 실험전우뇌  $\beta$ (고): 4/6

- F AND 실험전좌뇌  $\alpha$ (고) AND 실험전좌뇌  $\beta$ (고): 4/6

이것을 볼 때, 비교적 오류율이 높은 F형이라도 실험전 뇌파 수치가 높으면 오류율이 낮아짐을 잠정적으로 결론내릴 수 있다.

### 5.3.3 T형이 오류율이 (0.6보다) 더 낮은 경우

실험전  $\alpha$  좌우합계(저) AND 실험전  $\beta$  좌우합계(저)의 경우, F형은 오류율이 낮은 확률이 0(0/5)인데 반하여 T형은 0.85(6/7)이었다. 따라서 T형은 실험전 뇌파가 적으면 정확해진다는 가설을 세울 수 있다. 다음은 이를 뒷받침하는 일부 규칙들이다.

- T AND 실험전  $\beta$  좌우합계(저) AND 좌뇌  $\alpha$  전후증감(저): 4/4

- T AND 실험전  $\beta$  좌우합계(저) AND 실험중좌뇌  $\alpha$ (저): 5/5

- T AND 실험전좌뇌  $\beta$ (저) AND  $\beta$  전후증감좌우합계(저): 4/4

- T AND 실험전  $\alpha$  좌우합계(저) AND  $\beta$  전후증감좌우합계(저): 6/6

- T AND 실험중좌뇌  $\beta$ (저): 6/6

이들을 보면, T형이 뇌파증가량이 낮으면 정확도가 올라가지 않는가를 생각하게 된다. 그런데 여기에는 다음과 같은 반대 규칙들이 있다.

- T AND 좌뇌  $\beta$  전후증감(고): 5/9(무의미!)

- T AND 우뇌  $\alpha$  전후증감(고): 6/11(무의미!)

- T AND  $\beta$  전후증감좌우합계(고): 5/10(0.5)

또 일반적으로(T/F를 따지지 않고) 뇌파증가율이 낮아지는데도 오류율이 낮은 확률이 적은 다음과 같은 규칙들도 발견된다.

- 실험전우뇌  $\beta$ (저) AND 좌뇌  $\beta$  전후증감(저) AND 우뇌  $\beta$  전후증감(저): 1/5

- 실험전우뇌  $\beta$ (저) AND 좌뇌  $\alpha$  전후증감(저)

AND 좌뇌  $\beta$  전후증감(저): 1/5

### 5.3.4 잠정 결론

지금까지 관찰된 결과에 따라 세운 잠정 결론은 다음과 같다.

① T형의 경우 실험전 뇌파량이 적으면 정확도가 올라간다. (침착한 상태를 나타내는 것일지도 모른다.)

② T형과 뇌파증가율과는 관계가 없는 것 같다. 간혹 증가율이 높을 때 좋아 보이는 것은, 실험전 뇌파가 적어야 좋은 것의 영향이 아닐까 추측된다.

③ F형의 경우에는 실험전 뇌파량이 많아야 좋다. (침울하지 않고 활발한 상태를 나타내는 것일지도 모른다.)

④ F형과 뇌파증가율과는 관계가 없는 것 같다. 간혹 증가율이 낮을 때 좋아 보이는 것은, 실험전 뇌파가 많아야 좋은 것의 영향이 아닐까 추측된다.

그러나, 여기에 대해서도 반대증거들이 관측되고 있다. 따라서 앞으로 더 면밀한 분석이 필요하다.

## 6. 결론 및 앞으로의 연구 과제

지금까지 인지유형과 생리특성(뇌파)이 시계열 예측의 정확도에 미치는 영향을 데이터 마이닝을 통하여 분석해 보았다. 분석적인 사람이 직관적인 사람보다 더 정확한 경향을 찾을 수 있었다. 뇌파 증가율이 높거나 뇌파량이 적으면 분석적인 사람일 가능성이 많았다. 분석적인 사람은 실험 전에 뇌파량이 적을수록 더 정확해지며, 직관적인 사람은 실험 전에 뇌파량이 많을수록 더 정확해지는 것으로 현재까지 관측되었다. 그러나 반대 증거들도 관측되고 있으므로 더 면밀한 분석이 필요하다.

앞으로 새로운 데이터 항목을 유도해 내고 데이터 값의 구간도 더 세분화하며 실험도 더 많이 하여 더 정밀한 마이닝을 실행하여야 한다. 또 나타날 수 있는 규칙의 수가 막대하므로 고차원 인과관계를 이 연구에 알맞게 찾아낼 수 있는 자동화 도구가 필요하다.

### 참고 문헌

- [1] R. Agrawal, T. Imielinski, and A. Swami, "Database Mining: A Performance Perspective," IEEE TKDE, Vol. 5, No. 6, Dec. 1993.
- [2] A. Berson and S. J. Smith, "Data Warehousing, Data Mining, and OLAP," McGraw-Hill Pub., 1997.
- [3] M.-S. Chen, J. Han, and P. S. Yu, "Data

Mining: An Overview from a Database Perspective," IEEE TKDE, Vol. 8, No. 6, pp. 866-883, Dec. 1996.

[4] W. J. Frawley, G. Ipatetsky-Shapiro, and C. J. Matheus, "Knowledge Discovery in Databases: An Overview," Knowledge Discovery in Databases, AAAI/MIT Press, pp. 1-27, 1991.

[5] D. Heckerman, "Bayesian Networks for Data Mining," Data Mining and Knowledge Discovery, Kluwer Academy Pub., Vol. 1, No. 1, pp 79-119, 1997.

[6] M. Mehta, R. Agrawal, and J. Rissanen, "SLIQ: A Fast Scalable Classifier for Data Mining," EDBT, 1996.

[7] R. Ramakrishnan, "Database Management Systems," McGraw-Hill Pub., 1997.