논문 투고일:2016.10.17, 논문 심사일:2016.10.17-2016.12.21, 게재 확정일:2016.12.21

통계와 시각화를 결합한 데이터 분석: 예측모형 대한 시각화 검증

Data analysis by Integrating statistics and visualization: Visual verification for the prediction model

주저자

문성민 (Mun, Seong Min)

아주대학교 라이프미디어 협동과정 통합디자인연구실 연구원

낭떼르 대학교 언어과학, 박사과정

교신저자

이경원 (Lee, Kyung Won)

아주대학교 미디어학과

kwlee@ajou.ac.kr

본 연구는 2016학년도 아주대학교 BK사업단과 2015학년도 대한민국 교육부, 한국연구재단의 지원을 받아 수행된 연구임.
(NRF-2015S1A5B6037107)

목차

(요약)

예측 분석은 패턴인식(Pattern recognition) 혹은 기계학습(Machine learning)으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 사용자가 분석 과정에 개입하여 더 많은 정보를 얻어내기 위해서는 높은 통계적 지식수준이 요구된다. 또한 사용자는 분석 결과외의 다른 정보를 확인 할 수 없고 데이터의 특성 변화와 데이터 하나하나의 특징을 파악하기 힘들다는 단점이 있다. 본 연구는 이러한 예측분석의 단점을 보완하고자 통계적인 데이터 분석 방법과 시각화 분석 방법을 결합하여 데이터 분석을 진행하였으며 통계적인 분석 방법만을 진행 할 경우 발생하는 단점을 보완하고 데이터에서 더 많은 정보를 도출해 내기 위한 방법론을 제시 하고자하였다. 이를 위해 본 연구는 영화 리뷰에서 추출한 감정 어휘가 독립변인이고 영화의 흥행 값이 종속변인인 데이터를 예제 데이터로 활용하여 진행하였다. 본 연구의 연구 방법론을 적용하였을 때의 이점은 다음과 같다. 첫째, 의사결정나무 분석에서 제시된 분할 기준이 적용될 때 마다 변하는 데이터의 패턴을 파악할 수 있다. 둘째, 제시된 최종 예측모형에 포함된 데이터들의 특성을 확인 할 수 있다. 본 연구의 시사점은 예측모형의 단점을 보완하고 데이터로부터 더 많은 정보를 추출하기 위해 통계적인 데이터 분석과 시각적인 데이터 분석을 결합하여 시행하였다는 것이다. 통계적인 분석 방법을 통해 각 변수의 관계를 파악하고 높은 예측 값을 가지는 모형을 도출하였으며, 시각화 분석에서는 인터랙션 기능을 제공함으로서 통계적으로 제시된 예측모형을 검증하고 더 다양한 정보를 도출 할 수 있게 하였다.

(Abstract)

Predictive analysis is based on a probabilistic learning algorithm called pattern recognition or machine learning. Therefore, if users want to extract more information from the data, they are required high statistical knowledge. In addition, it is difficult to find out data pattern and characteristics of the data. This study conducted statistical data analyses and visual data analyses to supplement prediction analysis's weakness. Through this study, we could find some implications that haven't been found in the previous studies. First, we could find data pattern when adjust data selection according as splitting criteria for the decision tree method. Second, we could find what type of data included in the final prediction model. We found some implications that haven't been found in the previous studies from the results of statistical and visual analyses. In statistical analysis we found relation among the multivariable and deducted prediction model to predict high box office performance. In visualization analysis we proposed visual analysis method with various interactive functions. Finally through this study we verified final prediction model and suggested analysis method extract variety of information from the data.

(Keyword)

Data characteristics, Data pattern, Predictive model, Visualization

1. 서론

1-1. 연구배경 및 필요성

최근 정보통신의 발달과 함께 방대한 양의 데이터들이 생산되었으며 생산된 데이터를 활용, 분석하여 가치 있는 정보를 추출하고, 현상을 예측하는 예측분석의 활용이 중요해지고 있다. 예측분석이란 예측 모델링(Predictive modeling), 기계학습, 데이터 마이닝(Data mining) 등 과거의 데이터를 활용하여 미래의 행위를 예측하고 의사결정에 도움을 주는 통계적인 분석 방법이다.[1)]

예측분석을 활용하는 사례에 대한 일례로 2013년 경찰청에서 발표한 "지리정보 통합한 지리적 프로파일링 시스템 구축"에 따르면 최근 경찰청은 범죄수사의 범위를 줄여줄 지리적 프로파일링 시스템 개발을 위해 기존 발생한 범죄의 데이터를 통합 수집 및 분석을 수행하고 범죄의 가능성과 방향성을 기반으로 범죄발생 지역을 예측 한다고 한다. 또한 지리기반 데이터 시각화를 활용하여 예측 분석 결과를 범죄의 유형, 시간대에 따라 범죄다발지역과 위험도를 지도에 각기 다른 색으로 표시하여 수사 과정에 활용 한다고 한다.[2)] 이렇듯 데이터를 분석, 예측하여 실생활에 활용할 경우 낭비되는 많은 비용과 시간을 감소시키고 정확한 의사결정을 도울 수 있다. 예측분석 방법으로는 크게 회귀분석모형, 인공신경망분석, 사례기반추론, 유전자 알고리즘, 퍼지이론, 의사결정나무 분석 등이 있으며 본 연구에서는 의사결정나무 분석을 활용하여 연구를 진행 하고자 한다.[3)]

의사결정나무 분석은 다변량으로 이루어진 데이터를 분석하기에 적합한 통계적인 예측 분석 방법이며 패턴인식 혹은 기계학습으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 분석 결과의 정확도와 신뢰성이 높다. 하지만 분석에 사용되는 알고리즘이 복잡하고 많은 조건을 가정해야하는 어려운 분석 일수록 사용자가 분석 과정에서 더 많은 정보를 얻기 위해서는 높은 통계적 지식수준이 요구된다. 또한 사용자는 분석 결과외의 다른 정보를 확인 할 수 없고 데이터의 특성 변화와 데이터 하나하나의 특징을 파악하기 힘들다는 단점이 있다.[4)]

이러한 예측분석의 단점을 보완 할 수 있는 방법으로 최근에는 시각화 분석을 예측분석과 결합하여 분석을 진행함으로써 예측분석의 단점을 보완하고 사용자에게 더 많은 정보를 주기 위한 시도가 이뤄지고 있다. 2008년 발표된 Adam Perer외 1명의 연구에서는 데이터 분석 과정에서 통계적인 분석만을 수행 할 경우 데이터의 특이점이나 데이터 관계 내의 패턴을 파악하기 힘들지만 시각화 분석을 결합하여 사용할 경우 이러한 단점이 보완된다고 주장한 바 있다.[5)] 또한 2003년 발표된 Soon Tee Teoh외 1명의 연구에 따르면 의사결정나무 분석의 결과를 시각화 분석을 통해 확인하면 데이터의 군집화나 개별 데이터의 변화 패턴을 추가적으로 확인 할 수 있다고 주장하였다.[6)] 이러한 주장을 바탕으로 본 연구는 통계적인 데이터 분석 방법과 시각화 분석 방법을 결합하여 데이터 분석을 진행하고 통계적인 분석 방법만을 진행 할 경우 발생하는 단점을 보완하고 데이터에서 더 많은 정보를 도출해 내기 위한 방법론을 제시 하고자 한다.

예측분석의 단점을 보완하고 데이터로부터 더 많은 정보를 추출하기 위한 방법론을 제시하기 위해

---

1) David Lechevalier, Anantha Narayanan, Sudarsan Rachuri, "Towards a Domain-Specific Framework for Predictive Analytics in Manufacturing", 2014 IEEE International Conference on Big Data, p. 987, 2014.

2) 경찰청, "지리정보 통합한 지리적 프로파일링 시스템 구축 (GeoPros)", 2013 빅데이터 사례집, p.65, 2013.

3) Roiger, R., M. Heatz, "Data mining : A Tutorial Based Primer, Addison Wesley, 2003.

4) Soon Tee Teoh, KwanLiu Ma, "PaintingClass: Interactive Construction, Visualization and Exploration of Decision Trees", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, p. 668, 2003.

5) Adam Perer, Ben Shneiderman,"Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis", CHI 2008 Proceedings · Visual Synthesis, p. 265, 2008.

6) Soon Tee Teoh.

본 연구는 영화 리뷰에서 추출한 감정 어휘가 독립변인이고 영화의 흥행 값이 종속변인인 데이터를 예제 데이터로 활용하여 진행하였다. 본 연구는 통계적인 분석방법으로 영화의 흥행 값을 예측하기 위해서 의사결정나무 분석을 사용하였으며 최종 제시된 예측모형에 대한 검증하기 위해 시각화 분석을 사용하였다.

1-2. 연구의 목적 및 방법

본 연구는 통계적인 데이터 분석 방법과 시각적인 데이터 분석 방법을 결합하여 분석을 시행함으로써 예측분석의 단점을 보완하고 최종 제시된 예측모형에서 더 많은 정보를 추출하기 위한 방법론을 제시하는 것을 목적으로 진행하였다. 연구 목적을 달성하기 위한 연구 진행 과정은 다음과 같다.

첫째, 네이버 영화 평에서 리뷰의 개수가 1000개 이상인 672개의 영화에 대한 리뷰를 크롤링(Crawling)하고 감정 어휘 사전과 영화진흥위원회에서 제공하는 빈도 데이터를 활용하여 감정 어휘 데이터와 영화 흥행 값으로 구성된 예제 데이터를 생성하였다.

둘째, 높은 영화 흥행 예측 값을 도출하기 위해 전체 영화를 대상으로 의사결정나무 분석을 시행하였다.

셋째, 다양한 시각에서 데이터를 분석하기 위해 Parallel Coordinates 시각화를 제작하고 시각화 분석 방법을 활용하여 데이터를 분석 하였다.

넷째, Parallel Coordinates 시각화를 활용하여 의사결정나무 분석에서 제시된 최종 모형에 대한 검증을 수행하였다.

다섯째, 분석 결과를 해석하고 연구의 시사점과 연구의 한계, 향후 연구 방향을 제시하였다.

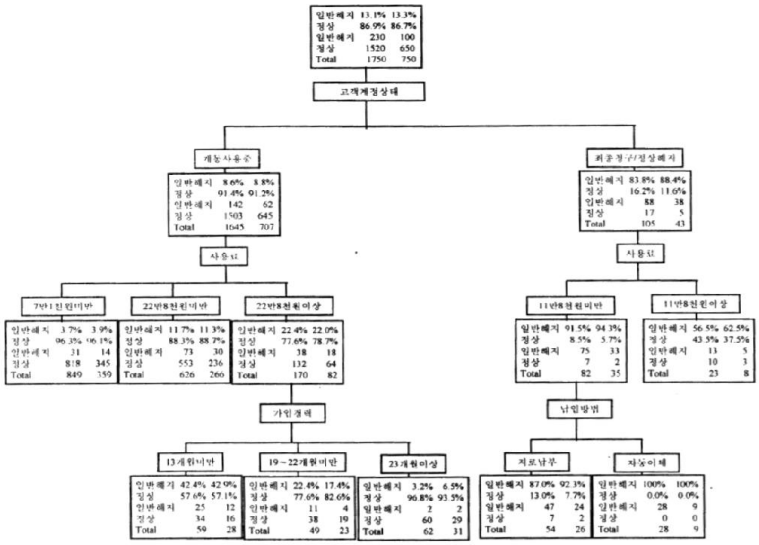
2. 이론 및 선행 연구의 고찰

2-1. 의사결정나무 분석 관련연구

예측분석 방법 중 하나인 의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 수행하는 분석방법이다. 또한 의사결정나무 분석은 다변량으로 이루어진 데이터 세트 내에서 목표가 되는 변수를 선정하고 대상이 되는 변수를 기준으로 높은 예측 값을 도출하기 위한 분할 기준과 분할 값을 도출하기 위해 사용 될 수 있다. 관련 연구로는 1998년 최종후, 서두성의 ‘의사결정나무를 이용한 개인휴대통신 해지자 분석’이라는 연구와 2014년 권영란, 김세영의 ‘의사결정나무 분석 기법을 이용한 중학생 인터넷게임중독의 보호요인 예측’ 등의 연구가 있다.

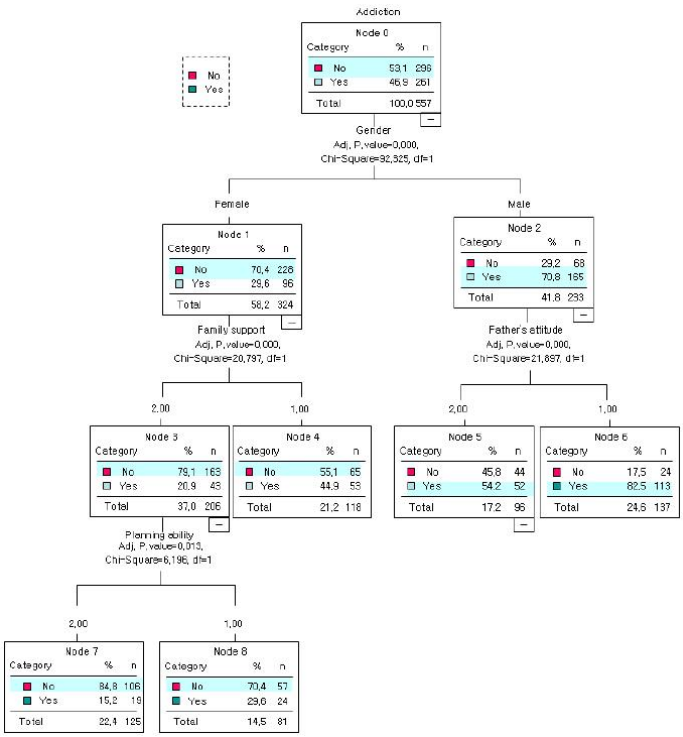
최종후, 서두성의 연구에서는 휴대전화 가입 고객의 해지를 결정하는 제일 중요한 변수는 고객계정 상태이며, 두 번째로는 최근 4개월간의 사용료, 세 번째로는 가입 경력과 납입 방법 등이 있다는 것을 도출 하였다. 또한 이중 가입고객의 고객계정 상태가 '최종청구/정상해지'인 경우 해지율이 83.8%, 88.4%로 높아진다는 것을 도출하였다.[7)]

7) 최종후, 서두성, "의사결정나무를 이용한 개인휴대통신 해지자 분석", 한국경영과학회, pp. 379, 1998.



〈그림 1〉 최종후 외 1명 연구의 〈그림1〉 "의사결정나무"

권영란, 김세영의 연구에서는 중학생의 인터넷게임중독에 영향을 미치는 보호요인으로 개인, 가족, 학교 관련 요인을 포괄적으로 규명하여 예측모형을 제시하였다. 분석결과 나무형태의 시각적 경로를 통하여 인터넷게임 일반 사용군에 포함될 확률이 가장 높은 경로는 여학생으로 가족 보호요인인 가족의 지지가 높고, 개인 보호요인인 계획성이 높은 경우인 것으로 도출되었으며 이에 비해 남학생의 경우에는 아버지의 태도가 엄격할수록 인터넷게임 일반 사용군에 포함될 확률이 높다는 결과를 제시하였다.[8)]



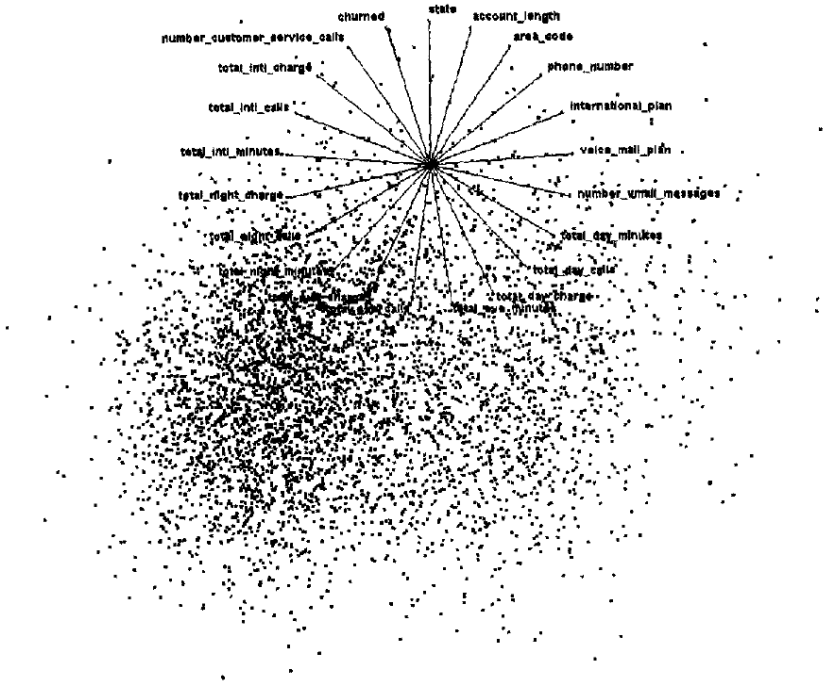
〈그림 2〉 권영란 외 1명 연구의 〈그림1〉 "The construction of decision tree."

8) 권영란, 김세영, "의사결정나무 분석 기법을 이용한 중학생 인터넷게임중독의 보호요인 예측", 정신간호학회지 13호, p. 19, 2014.

2-2. 시각화 분석 관련연구

의사결정나무 분석은 다변량으로 이루어진 데이터를 분석하기에 적합한 통계적인 분석 방법이다. 하지만 패턴인식 혹은 기계학습으로 불리는 확률적 학습 알고리즘을 기반으로 하는 통계적인 분석 일수록 데이터의 특성 변화를 파악하기 힘들다는 단점이 있으며 데이터 하나하나의 특성을 파악하지 못한다는 단점이 있다. 따라서 시각화 분야에서는 이러한 단점을 보완하기 위한 시도가 이루어지고 있다. 관련 연구로는 2001년 Eser Kandogan의 'Visualizing Multi-dimensional Clusters, Trends, and Outliers using Star Coordinates'등의 연구와 2003년 Soon Tee Teoh외 1명의 'Painting Class: Interactive Construction, Visualization and Exploration of Decision Trees'이라는 연구가 있다.

Eser Kandogan의 연구에서는 시각화 분석 방법 중 하나인 Star Coordinates활용하여 다변량의 데이터를 분석하는 방법을 제안하였다. Star Coordinates의 경우 이차원 공간상에서 하나의 위치 점을 기반으로 여러 변수 축들이 균등한 범위로 펼쳐 있다. 데이터는 펼쳐진 변수 축에서 높은 값을 가지는 방향으로 위치가 정해지는 방법으로 분류된다. Eser Kandogan는 본 연구에서 Star Coordinates를 활용하여 데이터를 분석 할 경우 특성이 비슷한 데이터를 군집화(Clustering)하는데 있어 유용하다는 점을 도출하였다.[9)]

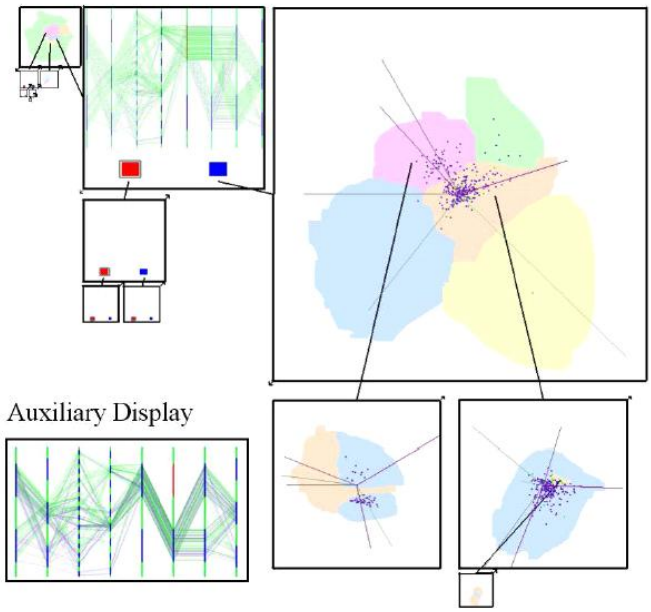


〈그림 3〉 Eser Kandogan 연구의 〈그림12〉 "Overview of 'churn' dataset, where churned customers are marked with blue (dark) color."

Soon Tee Teoh외 1명의 연구에서는 의사결정나무 분석 결과를 Parallel Coordinates와 Star Coordinates와 같은 시각화 분석 방법을 통해 나타냄으로써 데이터에서 발견할 수 있는 결과를 폭넓게 도출 하고자 하였다. 또한 이 두 시각화를 연결하여 사용하면 데이터 분류 과정과 데이터의 분류를 통합하여 확인 할 수 있다는 제안을 하였다.[10)]

9) E. Kandogan, "Visualizing Multi-Dimensional Clusters, Trends, and Outliers using Star Coordinates.", ACM SIGKDD '01, p. 113, 2001.

10) Soon Tee Teoh, KwanLiu Ma, "Painting Class: Interactive Construction, Visualization and Exploration of Decision Trees", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery



〈그림 4〉 Soon Tee Teoh 외 1명 연구의 〈그림5〉 "An auxiliary display is shown on the un-utilized space at the lower left of the display."

3. 예제 데이터 설명

본 연구는 아래와 같은 데이터 수집 과정을 통해 예제 데이터를 생성하였다.

첫째, 영화 리뷰 데이터를 수집을 위해 JAVA를 사용하여 국내에서 영화에 대한 의견 교류가 활발히 이루어지고 있는 네이버 영화 사이트에 대한 웹 크롤러를 제작하였다. 크롤러는 네이버 영화 홈페이지에서 특정 영화의 관람객 댓글과 리뷰들을 정제되지 않은 데이터 형태로 수집하도록 설계되었다.

둘째, 수집된 영화 데이터 중에서도 리뷰의 개수가 1000개 이상인 영화들만 다시 필터링하였고 최종적으로는 2289개의 영화 중 672개의 영화에 대한 리뷰 데이터가 수집되었다.

셋째, 선행 연구 중 문성민 외 2명의 연구(2015)를 참고하여 감정 어휘 사전을 구축하고 추출된 영화 리뷰 데이터에서 감정 어휘 값을 생성하였다. 이 과정을 통해 생성된 감정 어휘는 〈표 1〉과 같다.[11)]

〈표 1〉 문성민 외 2명의 연구, "최종 선정된 36개의 감정 어휘"

대표 감정 어휘	세부 감정 어휘
행복(Happy)	행복하다(Happy), 달콤하다(Sweet), 웃기다(Funny), 신나다(Exited), 기쁘다(Pleasant), 통쾌하다(Fantastic), 만족하다(Gratified), 재미있다(Enjoyable), 활기있다(Energetic)
놀라움(Surprise)	놀랍다(Surprised), 황홀하다(Ecstatic), 멋지다(Awesome), 훌륭하다(Wonderful), 대단하다(Great), 감동적이다(Touched), 인상깊다(Impressed)

and data mining, p. 670, 2003.

11) 문성민, 하효지, 이경원, "영화의 흥행 성과와 리뷰 감정 어휘와의 관계 분석", 디자인 융복합 학회, 제53(4)권, p.7, 2015.

지루함(Boring)	평온하다(Calm), 나른하다(Drowsy), 지루하다(Bored)
슬픔(Sad)	측은하다(Pitiful), 쓸쓸하다(Lonely), 애절한(Mournful), 슬프다(Sad), 비통하다(Heartbroken), 안타깝다(Unfortunate)
화남(Anger)	격분하다Outraged, 분노하다Furious
역겨운(Disgust)	불결하다(Ominous), 잔인하다(Cruel), 역겹다(Disgusted)
무서운(Fear)	공포스럽다(Scared), 등골이 서늘하다(Chilly), 섬뜩하다(Horrified), 무서워하다(Terrified), 오싹하다(Creepy), 무시무시하다(Fearsome)

넷째, 생성된 감정 어휘의 대표 감정 어휘와 영화진흥위원회에서 추출한 각 영화의 객관적인 데이터를 병합하여 최종적인 데이터를 생성하였다.[12)]

최종적으로 생성된 독립변수 데이터는 7개의 대표 감정 어휘('Happy', 'Surprise', 'Boring', 'Sad', 'Anger', 'Disgust', 'Fear')와 영화 티켓 판매액, 영화 관람 관객 수, 상영 스크린 수, 한 스크린 당 영화 관람 관객 수, 영화의 장르, 영화의 영문 이름이다. 또한 1983년 리트만(Litman)의 연구를 참고하여 누적 관객 수를 상영 스크린 수로 나누어 한 스크린에서의 누적 관객 수를 본 연구의 종속변수인 영화 흥행 값으로 사용하였다. 최종 생성된 예제 데이터에 대한 기술통계량은 〈표 2〉와 같다.

〈표 2〉 최종 생성된 예제 데이터에 대한 기술통계량

변수명	영문명	최댓값	평균	최솟값
판매액	Sales	1.280e+11	1.019e+10	1.670e+04
관객 수	Attendance	13624328	1469038	35
개봉 스크린 수	Screen	1409.0	330.7	1.0
흥행 값 (평균 관객 수)	Normal_attendance	33590	4018	35
기쁨	Happy	0.0400	0.3073	0.7400
놀라움	Surprise	0.0600	0.2625	0.6500
지루함	Boring	0.02000	0.09126	0.32000
슬픔	Sad	0.0300	0.1465	0.5400
화남	Anger	0.01000	0.06247	0.28000
역겨움	Disgust	0.00000	0.04314	0.37000
무서움	Fear	0.00000	0.08708	0.65000

4. 통계를 활용한 예측 분석

데이터를 활용한 연구들은 주로 탐색적 연구(Exploratory research), 기술적 연구(Descriptive research), 인과관계 연구(Causal research)를 사용하며 이러한 연구는 통계학에 기반을 두고 있다. 통계학은 사회와 사회 구성원에게서 수집된 양적/질적 자료를 기술하고 해석하기 위한 방법을 연구하는 것으로 신뢰도 95%에서 기각역을 α 혹은 p(Probability)〈.05의 수준으로 정하고 통계분석 결과가 이를 만족하면 유의미한 결과라고 해석한다.[13)] 하지만 분석에 사용되는 알고리즘이 복잡하고 많은 조건을 가정해야하는 어려운 분석 일수록 사용자가 분석 과정에 개입하기는 많은 지식수준이 요구된다. 따라서 사용자는 분석 결과외의 다른 정보를 확인 할 수 없기 때문에 데이터의 특성 변화와 데이터 하나하나의 특징을 파악하기 힘들다는 단점이 있다.[14)] 최근에는 이러한 단점을 보완하고 데

12) 영화진흥위원회, http://www.kofic.or.kr/

13) DeGroot, Schervish, "Definition of a Statistic". Probability and Statistics Third Edition Addison Wesley, pp.370-371, 2002.

14) Soon Tee Teoh.

이터로부터 더 많은 정보를 얻어내기 위해 시각화 분석을 결합하여 분석을 진행하고 있다.[15] 시각화 분석이란 데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 분석 방법으로써 연결과 그룹화를 통한 데이터 요약, 색, 모양 등 미적요소를 활용한 데이터의 특성 표현 등 다양한 방법으로 사용자의 이해를 돕는 분석이다.[16] 본 연구는 통계적 분석과 시각화 분석을 결합하여 연구를 진행하였고 통계적 분석을 수행 후 분석 결과에 대해 시각화 분석을 수행함으로써 이를 검증하였다.

4-1. 의사결정나무 분석의 정의

의사결정나무는 의사결정규칙을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류하거나 예측을 수행하는 분석방법이다. 이는 방대한 양의 데이터베이스에서 연구자가 원하는 목표변수 값에 도달하기 위해 영향을 미치는 변수들을 도출해내고 최적의 분리 기준을 찾아 의사결정에 도움을 주는 일련의 과정이라고도 이야기 할 수 있다.[17]

의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용 될 수 있으나 분석의 정확도보다는 분석 과정의 설명이 필요한 경우에 더 유용하게 사용된다. 의사결정나무 분석이 활용될 수 있는 응용분야는 〈표 3〉과 같다.

〈표 3〉 의사결정나무 분석 응용분야

용도	설명
세분화(Segmentation)	관측개체를 비슷한 특성을 갖는 몇 개의 그룹으로 분할하여 각 그룹별 특성을 발견 하고자 하는 경우
분류(Classification)	여러 예측변수(Predicated variable)에 근거하여 목표변수(Target variable)의 범주를 몇 개의 등급으로 분류 하고자 하는 경우
예측(Prediction)	자료로부터 규칙을 찾아내고 이를 이용하여 미래의 사건을 예측 하고자 하는 경우
차원축소 및 변수선택(Data reduction and variable screening)	매우 많은 수의 예측변수 중에서 목표변수에 큰 영향을 미치는 변수들을 골라 내고자 하는 경우
교호작용효과의 파악(Interaction effect identification)	여러 개의 예측변수들이 결합하여 목표변수에 작용하는 교호작용을 파악 하고자 하는 경우
범주의 병합 또는 연속형 변수의 이산화(Category merging and discretizing continuous variable)	범주 형 목표변수의 범주를 소수의 몇 개로 병합하거나, 연속형 목표변수를 몇 개의 등급으로 범주화 하고자 하는 경우

의사결정나무 분석은 목표변수, 예측변수, 분리기준, 분리개수에 따라 크게 CHAID(CHi-squared Automatic Interaction Detection), Exhaustive CHAID, CART(Classification And Regression

15) Adam Perer.

16) Pak Chung Wong, J. Thomas, "Visual Analytics", IEEE Computer Graphics and Applications Volume 24 Issue 5, pp. 20, 2004.

17) Soon Tee Teoh.

Trees), QUEST(Quick Unbiased Efficient Statistical Tree)로 나누어진다. 언급된 네 가지 분석 방법에 대한 설명은 〈표 4〉와 같다.

〈표 4〉 의사결정나무 분석의 종류

	CHAID	Exhaustive CHAID	CART	QUEST
목표변수	질적변수, 양적변수	질적변수, 양적변수	질적변수, 양적변수	명목형 질적변수
예측변수	질적변수, 양적변수	질적변수	질적변수, 양적변수	질적변수, 양적변수
분리기준	F검정, 카이제곱통계량	F검정, 카이제곱통계량	지니계수감소	F검정, 카이제곱통계량
분리개수	다지분리	다지분리	이지분리	이지분리

본 연구에서 분석에 사용될 영화 흥행 값과 7가지 대표 감정 어휘 값의 경우 연속형으로 이루어진 데이터 세트이다. 따라서 네 가지의 분석 방법 중 목표변수(종속변수)와 예측변수(독립변수)로 연속형 데이터(양적변수)를 다루고, 분리개수가 이지분리(Binary split)를 따르는 CART 분석 방법을 사용하였다. CART는 종속변수에 대하여 가능한 많은 동질적인 데이터가 같은 그룹에 속하도록 노드를 수정하는 방법을 사용하는데 분할 규칙으로는 데이터 내에서 가능한 모든 분할 규칙 중에서 불순도 값이 가장 최소가 되는 것을 따른다. 또한 불순도 함수로 지니 지수(범주 형 목표변수인 경우 적용) 또는 분산의 감소량(연속형 목표변수인 경우 적용)을 이용하여 이지분리를 수행하는 알고리즘이다. 가장 널리 사용되는 의사결정나무 알고리즘으로 개별 입력변수 뿐만 아니라 입력변수들의 선형 결합들 중에서 최적의 분리를 찾을 수도 있다.

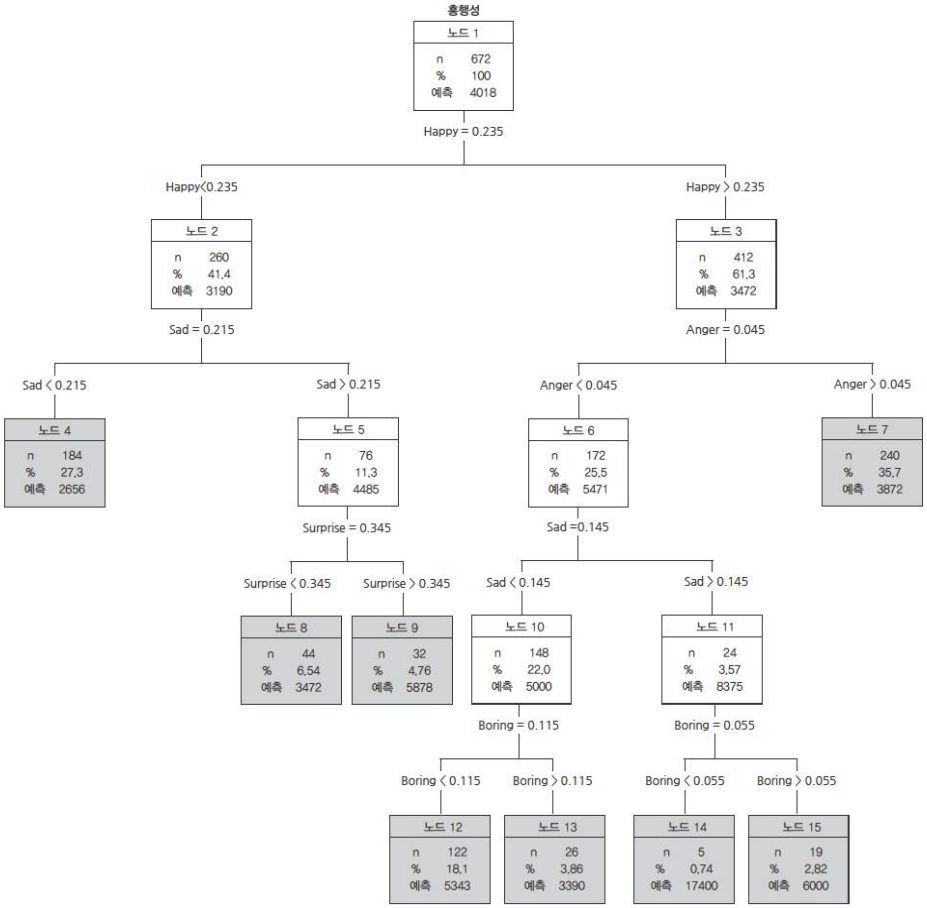
4-2. 의사결정나무 분석 결과

전체 영화 데이터 집단에 대한 의사결정나무 분석을 수행 하기에 앞서 목표가 되는 종속변수와 분할 기준으로 작용을 할 독립변수와 종속변수에 대한 기술 통계 값은 〈표 5〉와 같다.

〈표 5〉 변수별 기술통계량

변수구분	변수명	평균	최댓값	최솟값
종속변수	흥행 값 (Normal_attendance)	4018	33590	35
독립변수	Happy	0.3073	0.74	0.04
	Surprise	0.2625	0.65	0.06
	Boring	0.0912	0.32	0.02
	Sad	0.1465	0.54	0.03
	Anger	0.0624	0.28	0.01
	Disgust	0.0431	0.37	0.00
	Fear	0.0870	0.65	0.00

분석에 사용된 전체 영화 데이터 집단에 대한 기술 통계량 값을 보면 흥행 값은 종속변수로써 평균값이 4018이며 7개의 대표 감정 어휘 값들이 독립변수로 사용되었고 Happy와 Surprise의 경우 감정어의 평균값이 다른 감정 어휘보다 높은 것을 확인 할 수 있다.



〈그림 5〉 전체영화에 대한 의사결정나무 분석 결과

전체 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 영화 흥행 값이 가장 높다고 예측된 집단에 대한 해석은 다음과 같다. 영화 흥행 값이 17400이 되기 위해서는 Happy 0.235를 기준으로 최초 분리 되어야 하며 Happy가 0.235이상 일 경우 Anger 0.045를 기준으로 다시 분리된다. Anger가 0.045이하 일 경우 다시 Sad 0.145를 기준으로 분리되며 Sad가 0.145 이상 일 경우 마지막으로 Boring 0.055를 기준으로 분리되며 Boring이 0.055이하 일 때의 집단(N=5)에 대한 영화 흥행의 예측 값은 17400으로 높게 분류된다. 분석된 의사결정 나무분석 결과는 〈그림 5〉와 같다.

의사결정 나무 분석의 결과는 자료의 분류가 얼마나 잘되었는지 한눈에 표현하는 이익도표를 통해 더 자세히 확인 할 수 있다. 전체 영화에 대한 이익도표 값은 〈표 6〉과 같다.

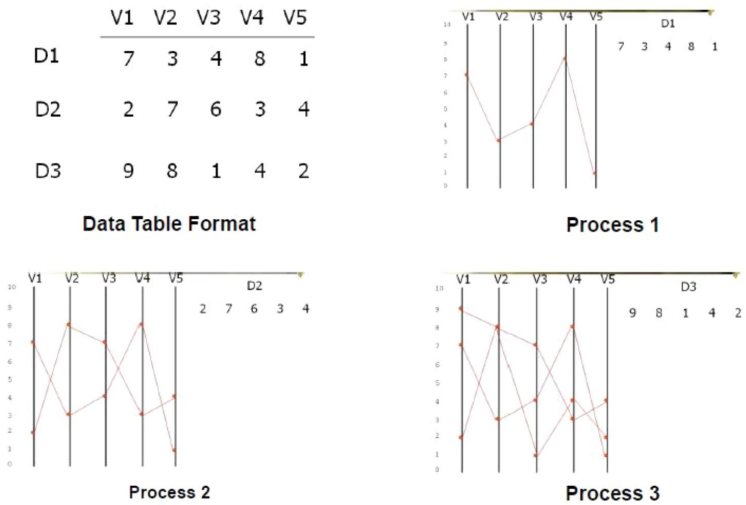
〈표 6〉 전체 영화에 대한 이익도표

노드번호	개수(N)	비율(%)	영화 흥행 예측 값
14	5	0.74	17400
15	19	2.82	6000
9	32	4.76	5878
12	122	18.1	5343
7	240	35.7	3872
8	44	6.54	3472
13	26	3.86	3390
4	184	27.3	2656

5. 시각화 구축 및 검증

5-1. Parallel Coordinates의 개념

본 연구의 데이터처럼 다변량으로 되어 있는 데이터를 분석하기 위해 여러 시각화 분석 방법 중 Parallel Coordinates를 사용하는 것이 적절하다. Parallel Coordinates 시각화 분석 방법은 N차원 공간 안의 점들의 집합을 보여주기 위한 방법으로 일반적으로 수직의 형태이며 N개의 등 간격 평행 라인으로 이루어져 있다. 또한 시계열 데이터 시각화에도 밀접한 관계가 있으며 데이터 내 변수간의 관계를 파악하는데 용이하다.[18] 이 방법은 1985년 Inselberg. A. 가 구체적으로 제안하였고 최근까지 다양한 학문 영역에서 사용되고 있다. Inselberg. A.의 연구에 따르면 Parallel Coordinates는 각 변수가 대부분 라인이 평행일 때 두 차원 사이에 유사한 관계가 형성된다고 해석할 수 있으며, 대부분의 라인이 교차할 때는 상이한 관계가 형성된다고 해석한다.[19]



〈그림 6〉 분포에 따른 Parallel Coordinates

또한 본 연구에서는 다변량 데이터의 분석 및 통계분석 결과에 대한 검증을 실시하기 위해 기존의 Parallel Coordinates 시각화 방법에 분석 목적에 부합하는 여러 기능을 추가하였다. 추가된 기능으로는 선택된 데이터의 평균값을 나타내는 기능, 영화의 장르를 선택하는 기능, 축을 변경하는 기능, 축을 제거하는 기능, 하나의 영화를 선택하여 데이터의 특징을 확인하는 기능, 영화의 제목 명으로 데이터를 검색하는 기능, 선택되지 않은 영화를 표현하는 기능 등 분석에 필요한 다양한 인터랙션 기능들이 있다.

5-2. 구축된 Parallel Coordinates의 기능

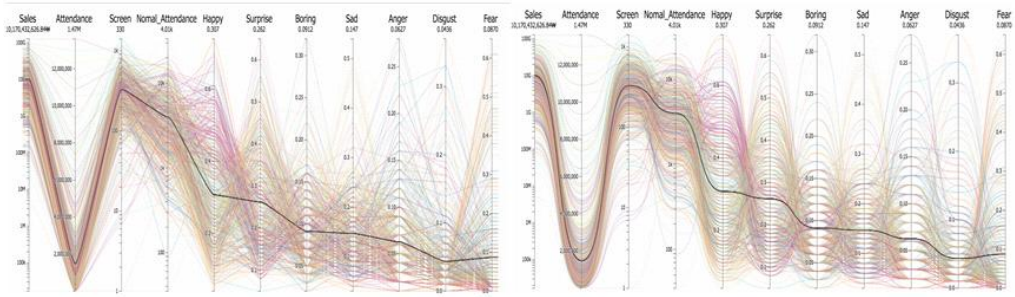
본 연구에서는 다변량으로 이루어진 데이터를 분석하고 통계분석 결과를 검증하기 위해 기존의 Parallel Coordinates 시각화 방법에 다양한 기능을 추가하였다. 해당 시각화는 http://202.30.24.167:8080/parallel.html에서 사용할 수 있으며, 연구에 사용된 Parallel Coordinates 시각화의 기능은 다음과 같다.

18) Rick Walker, Philip A. Legg, Serban Pop, Zhao Geng, Robert S. Laramee, Jonathan C. Roberts, "Force-Directed Parallel Coordinates", 17th International Conference on Information Visualisation, p.39, 2013.

19) Inselberg, A, The plane with Parallel Coordinates, The Visual Computer, p.79, 1985.

5-2-1. 번들링(Bundling)

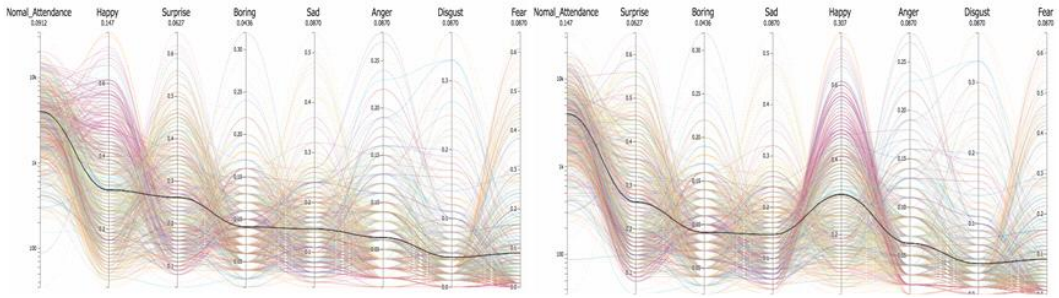
Parallel Coordinates는 일반적으로 데이터의 연결 표현을 직선으로 표현한다. 하지만 직선으로 데이터 연결을 표현하면 데이터의 양이 많을 때 축이 보이지 않고 데이터 사이의 패턴 또한 발견하기 어렵다. 따라서 본 연구는 〈그림 7〉의 오른쪽과 같이 번들링 기능을 추가 하였다. 번들링 기능을 통해 데이터 사이의 연결을 표현함으로써 데이터들이 군집화 되는 경향을 쉽게 확인 할 수 있다.



〈그림 7〉 (왼쪽) 직선으로 표현된 Parallel Coordinates (오른쪽) 번들링으로 표현된 Parallel Coordinates

5-2-2. 축(Axes)

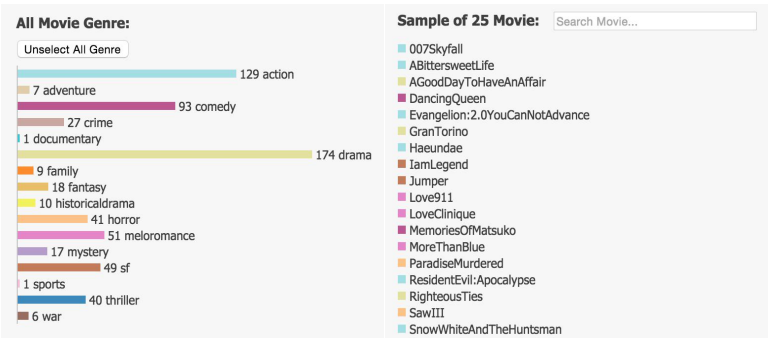
본 연구는 〈그림 8〉과 같이 Parallel Coordinates에 데이터 변수 축을 삭제 , 혹은 축의 순서를 이동시키는 기능을 추가하여 분석을 용이하게 하였다. 이를 통해 중요 변수별로 축을 나열 할 수 있으며 불필요한 변수를 삭제 할 수도 있다.



〈그림 8〉 (왼쪽) 변경 전 데이터 축의 순서 (오른쪽) Happy의 데이터 변수 변경 후 축의 순서

5-2-3. 색상(Colour)

장르별로 비교 분석을 하기 위해서는 영화의 장르별로 구분 할 수 있는 기능이 필요하다. 이를 위해 본 연구는 〈그림 9〉와 같이 영화의 장르에 따라 색상을 다르게 지정하여 사용자가 데이터를 쉽게 구분 할 수 있도록 하였다.

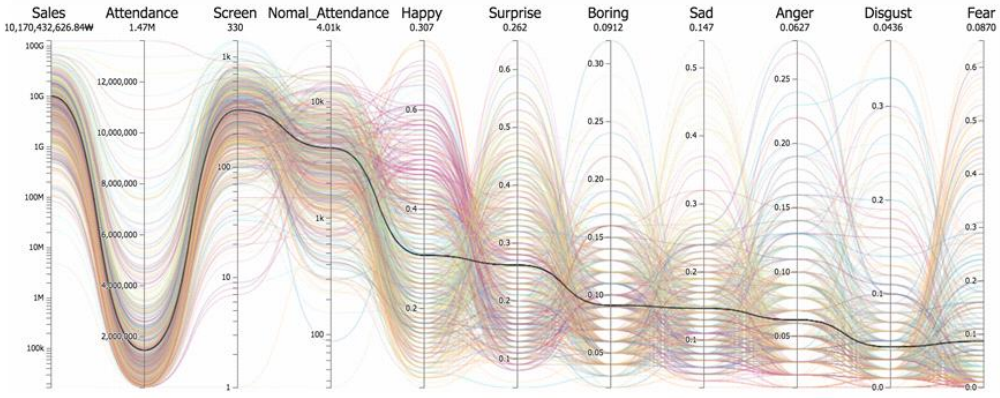


〈그림 9〉 영화 장르에 따라 지정된 색상

5-2-4. 기술 통계(Descriptive Statistic)

일반적인 Parallel Coordinates는 데이터 패턴, 데이터 변수 축에서 발생하는 군집화, 변수 축 사이의 직선 기울기 등 시각적으로 확인이 가능한 부분만으로 해석을 해야 한다. 본 연구는 〈그림 10〉과 〈그림 11〉과 같이 선택된 데이터의 평균선, 선택된 데이터 변수 축의 평균값, 선택된 영화 수의 합계를 내타내는 기능을 추가하였다. 이를 통해 사용자는 시각화로부터 더 다양한 데이터 정보를 얻을 수 있다.

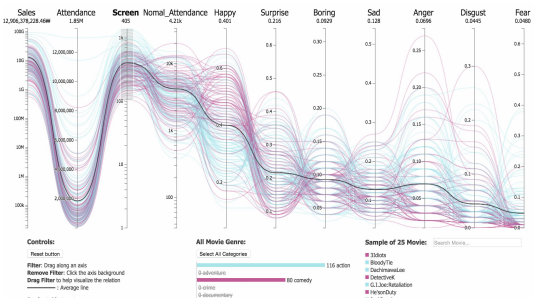
〈그림 10〉 선택된 데이터 변수들의 평균값과 영화 수의 합계



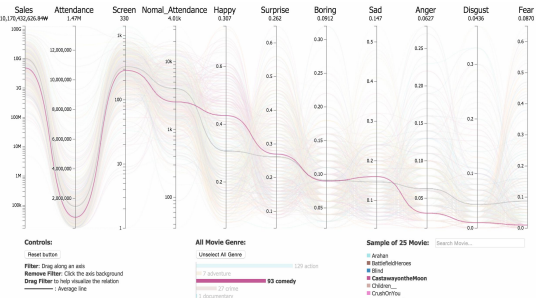
〈그림 11〉 선택된 데이터 변수들의 평균값과 평균선 (굵은 line 그래프)

5-2-5. 데이터 선택(Data Selection)

장르 별 데이터의 패턴 비교, 분포 확인, 조건에 따른 패턴 변화 등을 확인하기 위해서는 데이터를 선택하고 지정하는 기능이 필요하다. 본 연구에서는 분석의 용이성을 높이기 위해 〈그림 12〉와 〈그림 13〉과 같이 장르 선택 기능, 영화 검색 기능, 조건에 따른 데이터 필터링 기능, 하이라이트 기능 등을 추가하여 분석을 용이하게 하였다.



〈그림 12〉 데이터 선택: 장르가 액션 & 코미디이고 상영 스크린 수가 100개 이상인 영화

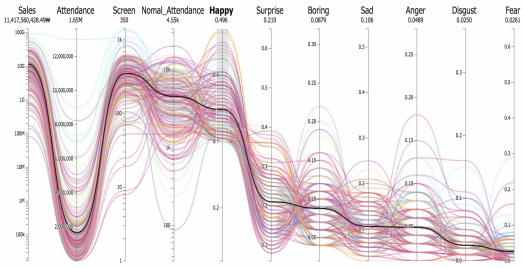


〈그림 13〉 2009년 개봉작 김씨 표류기(Castaway on the Moon)에 대한 하이라이트 View

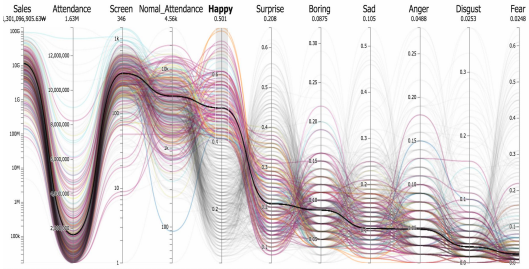
5-2-6. 제거된 데이터 표현

선행된 통계적인 분석 방법 중 의사결정나무 분석 과정을 Parallel Coordinates를 통해 검증하기 위해서는 선택되지 않은 데이터를 얕은 배경으로 표현함으로써 제거된 데이터의 규모를 보여주는 기능이 요구된다. 따라서 기존의 Parallel Coordinates기능에 선택되지 않은 데이터를 표현하는 방법을

추가하였다. 추가된 시각화는 〈그림 14〉, 〈그림 15〉와 같다.



〈그림 14〉 삭제된 데이터 표현 기능 추가 전



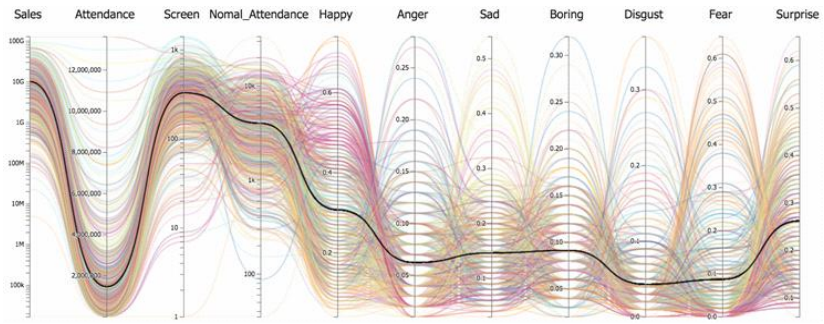
〈그림 15〉 삭제된 데이터 표현 기능 추가 후

5-3. 통계분석 결과에 대한 시각화 검증

본 장에서는 선행된 의사결정나무 분석 예측모형에 대해 시각화 분석 방법을 활용하여 검증 하고자 한다. 의사결정나무 분석을 활용하여 도출된 예측모형은 비슷한 감정을 느끼는 장르별로 군집 화된 집단과 전체 영화집단에 대하여 어떠한 감정이 느껴질 경우 흥행의 예측 값이 최고가 될 수 있는지를 제안하는데 매우 유용하게 활용될 수 있다. 하지만 의사결정나무 분석의 경우 패턴인식 혹은 기계학습으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 분석된 결과 외에는 일반적인 사용자가 유동적으로 분석 과정을 볼 수 없다는 단점이 있다.[20] 따라서 본 장에서는 각 집단에 따라 높게 예측된 노드에 대한 분류기준을 개발된 Parallel Coordinates 시각화 방법을 통하여 검증하고 시각화 분석 방법을 결합하여 사용자가 유동적으로 분석 과정에 참여하는 방법을 제안 하고자 한다.

5-3-1. 의사결정나무 분석 결과에 대한 시각화 검증

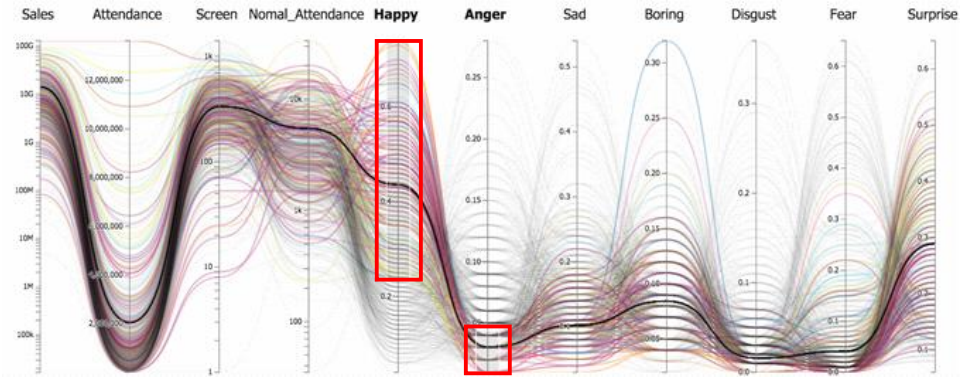
전체 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 의사결정나무 분석을 통해 영화 흥행의 예측 값이 높게 측정된 노드 14에 해당하는 집단의 분할 규칙은 Happy 〉 0.235 & Anger 〈 0.045 & Sad 〉 0.145 & Boring 〈 0.055의 순서로 4번 분할 된 것을 확인 할 수 있는데 본 연구에서는 최대의 흥행 값이 예측된 노드 14에 대한 데이터 분할 과정을 Parallel Coordinates시각화 분석 방법을 통해 검증 하고자 한다. 노드 14에 대해서 분할 기준이 적용되지 않은 전체 값은 〈그림 16〉과 같다.



〈그림 16〉 분할 기준이 적용되기 전의 시각화

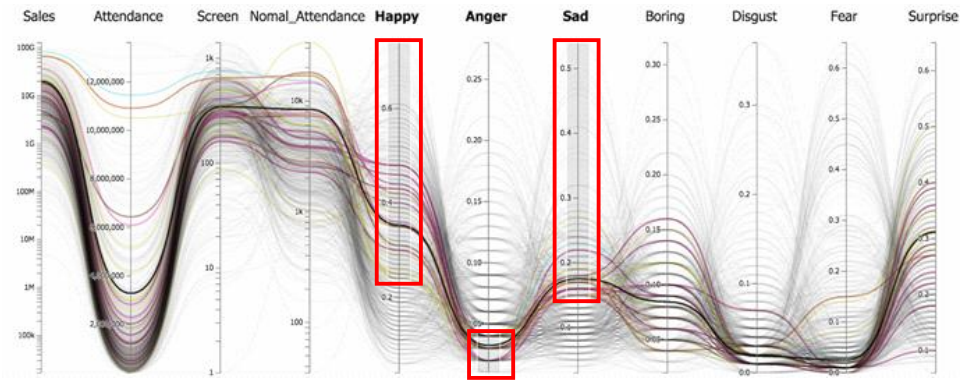
다음으로 Happy의 값이 0.235이상 일 때와 아닐 때로 최초 분할되었으며 Happy의 값이 0.235이상인 노드에 대한 시각화 결과는 〈그림 17〉과 같다.

20) Soon Tee Teoh, KwanLiu Ma. p. 667, 2003.



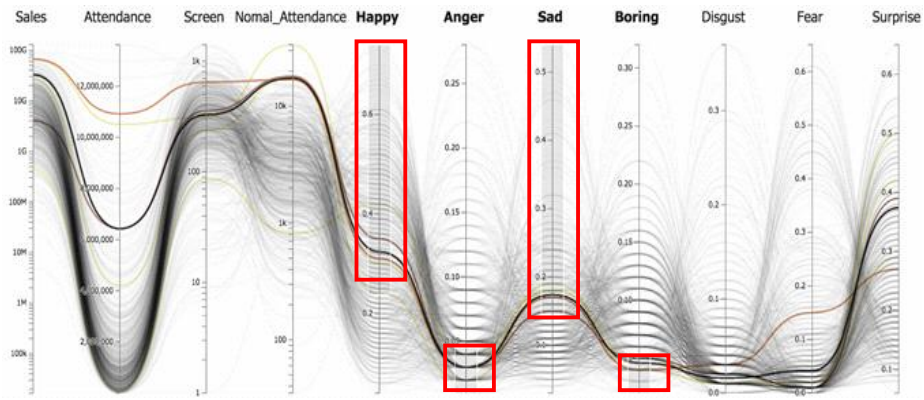
〈그림 17〉 Happy 〉 0.235 & Anger 〈 0.045이 적용된 시각화 결과

Happy를 기준으로 0.235이상인 데이터를 선택하였을 때 Happy의 값이 대체적으로 낮았던 드라마 장르의 영화와 호러 장르의 영화가 대폭 감소하였다. 드라마의 경우 174에서 79로, 호러의 경우 41에서 2로 감소한 것을 확인하였다. 노드 14에 대해 두 번째로 적용된 분할 기준은 Anger이며 Anger값이 0.045이상 일 때와 아닐 때로 분류된 시각화 결과는 〈그림 18〉과 같다.



〈그림 18〉 Happy 〉 0.235 & Anger 〈 0.045 & Sad 〉 0.145이 적용된 시각화 결과

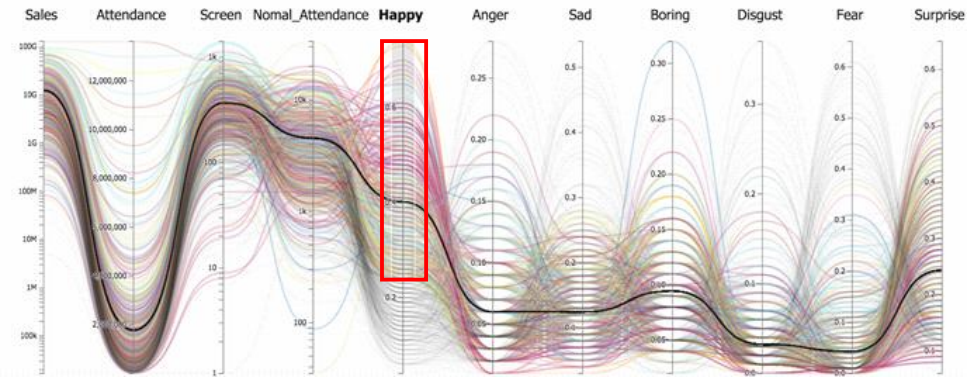
Happy는 0.235이상이고 Anger는 0.045이하인 데이터를 선택하였을 때 액션, 판타지, SF장르의 영화가 대폭 감소하였다. 액션의 경우 107에서 24로, 판타지의 경우 18에서 2로, SF의 경우 31에서 5로 감소한 것을 확인하였다. 노드 14에 대해 세 번째로 적용된 분할 기준은 Sad이며 Sad값이 0.145이상 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각화 결과는 〈그림 19〉와 같다. Happy는 0.235이상, Anger는 0.045이하, Sad는 0.145이상인 데이터를 선택하였을 때 코미디와 드라마장르가 각각 8편과 10편으로 선택된 23편의 영화 중 가장 많은 비율을 차지하였다. 마지막으로 적용된 분할 기준은 Boring이며 Boring값이 0.055이하 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각



〈그림 19〉 노드 14 에 대한 최종 Parallel Coordinates

화 결과는 〈그림 20〉, 〈그림 21〉과 같다.

〈그림 20〉 노드 14 에 최종 포함된 영화 정보



〈그림 21〉 Happy 〉 0.235이 적용된 시각화 결과

의사결정나무 분석을 통해 영화 흥행의 예측 값이 높게 측정된 노드 14에 대해서 Parallel Coordinates시각화 분석 방법을 사용하여 분석 한 결과, 노드 14 집단에 최종 포함된 영화는 King And The Clown, Malaton, NANA, The Host, Welcome to Dongmakgol이었으며 영화의 장르는 Drama 3, SF 1, War 1로 드라마 장르가 많이 포함된 것을 확인 할 수 있었다. 또한 최종 선택된 영화에 대한 선택기능을 통해 최종노드에 포함된 데이터들도 서로 다른 특성을 가지고 있는 것을 발견할 수 있었다.

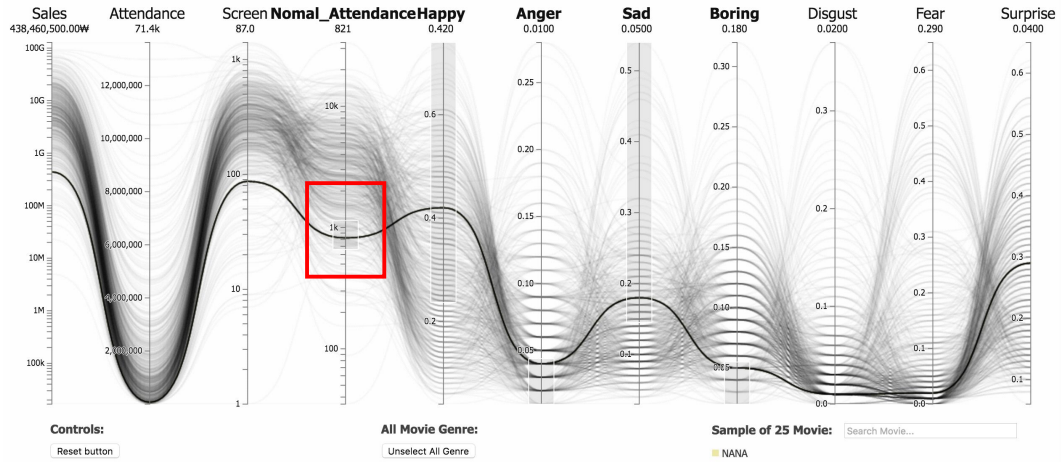
5-3-1. 의사결정나무 분석 결과에 대한 시각화 검증

본 장에서는 선행된 의사결정나무 분석을 통해 생성된 최종 예측모형을 Parallel Coordinates 시각화 방법을 통하여 검증하고 시각화 분석 방법을 결합하여 사용자가 유동적으로 분석 과정에 참여하는 방법을 제안하였다. 분석 과정에 대해서 Parallel Coordinates를 활용한 검증이 수행되면 분할 기준에 따른 데이터의 특성 변화를 파악 할 수 있으며 통계적인 분석 방법에서 발견하지 못한 결과를 도출해 낼 수 있다. 시각화 분석 방법을 활용해 의사결정분석 결과를 검증하였을 때 사용자가 추가로 얻을 수 있는 결과는 크게 두 개로 나누어 볼 수 있다.

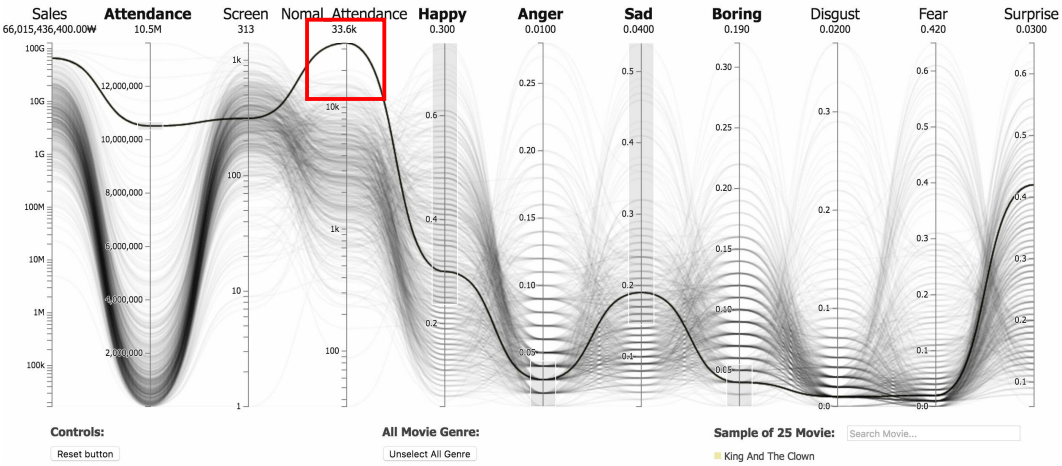
첫째, 분할 기준이 적용될 때 마다 변하는 데이터의 패턴을 파악할 수 있다. 예를 들어, 첫 번째 분할 기준으로 Happy의 값이 0.235이상인 노드가 선택 되었을 때 Happy의 값이 낮았던 드라마 장르의 영화가 174에서 79로 대폭 감소한 결과를 확인하였다.

둘째, 최종노드에 포함된 데이터들도 서로 상이한 특성을 지니고 있다는 것을 확인 할 수 있다. 〈그

림 22〉와 〈그림 23〉을 를 통해 노드 14에 포함된 영화 중에서도 흥행 값이 가장 높았던 영화는 King And The Crown으로 흥행 값은 33,600이었고 흥행 값이 가장 낮았던 영화는 NANA로 흥행 값이 821이라는 것을 확인 할 수 있다.



〈그림 22〉 노드 14 에 최종 포함된 NANA에 대한 시각화 결과



〈그림 23〉 노드 14 에 최종 포함된 King And The Crown에 대한 시각화 결과

6. 결론

산업의 성장과 함께 방대한 양의 데이터들이 생산되었으며 생산된 데이터를 활용, 분석하여 가치 있는 정보를 추출하고, 현상을 예측하는 예측분석의 활용이 중요해지고 있다. 예측분석은 패턴인식 혹은 기계학습으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 분석 결과의 정확도와 신뢰성이 높다. 하지만 분석에 사용되는 알고리즘이 복잡하고 많은 조건을 가정해야하기 때문에 사용자가 분석 과정에서 다양한 정보를 얻기 위해서는 많은 통계적 지식이 요구된다. 따라서 사용자는 분석 결과 외의 다른 정보를 확인 할 수 없고 데이터의 특성 변화와 데이터 하나하나의 특징을 파악하기 힘들다는 단점이 있다. 본 연구는 이러한 단점을 보완하고 데이터로부터 더 다양한 정보를 추출하기 위해 통계적인 데이터 분석 방법과 시각화 분석 방법을 결합하여 분석을 수행하였다. 분석에는 영화의 흥행 값과 영화 리뷰에서 추출한 감정 어휘 값으로 이루어진 데이터가 활용되었다. 영화의 흥행 값을 예측하기 위해 예측분석의 한 종류인 의사결정나무 분석을 수행하고 다양한 시각화 분석 기법 중에서 Parallel Coordinates를 활용하여 예측모형을 검증하였다. 본 연구의 시사점은 다

음과 같다.

첫째, Parallel Coordinates 시각화 분석을 활용하면 의사결정 나무 분석에서 제시된 예측모형의 분할 기준이 적용될 때 마다 변하는 데이터의 패턴을 파악할 수 있다. 예를 들어, 첫 번째 분할 기준으로 Happy의 값이 0.235이상인 노드가 선택 되었을 때 Happy의 값이 낮았던 드라마 장르의 영화가 174에서79로 대폭 감소한 결과를 확인하였다. 이는 Parallel Coordinates의 기능 중 조건에 따라 데이터를 선택하는 기능과 장르에 따라 색상을 달리 부여하는 기능을 활용한 결과로써 예측분석의 분할 기준을 시각화를 활용하여 분석함으로써 도출된 결과라고 할 수 있다. 이를 본 연구에서 사용된 데이터가 아닌 일반적인 데이터에 빗대어 해석하면 데이터가 지니는 인구통계학적 특성에 따라서 데이터는 서로 상이한 특성을 지니고 있으며 적용되는 분할 기준에 따라 선택, 제거되는 데이터의 특성도 변화한다고 할 수 있다.

둘째, 최종노드에 포함된 데이터들도 서로 상이한 특성을 지니고 있다는 것을 확인 할 수 있다. 의사결정나무 분석 결과에서 가장 높은 흥행 예측 값을 보인 노드 14에 포함되는 데이터들의 특성을 확인하여 본 결과 노드 14에 포함된 영화 중에서도 흥행 값이 가장 높았던 영화는 King And The Crown으로 흥행 값은 33,600이었고 흥행 값이 가장 낮았던 영화는 NANA로 흥행 값이 821이었다. 이러한 결과를 통해 예측분석으로 도출된 최종 모형 내에 포함된 데이터들도 사이에도 서로 상이한 특성이 존재한다는 것을 확인하였으며 시각화 분석을 사용할 경우 이러한 관계를 더 잘 확인할 수 있었다.

본 연구의 시사점은 예측모형의 단점을 보완하고 데이터로부터 더 많은 정보를 추출하기 위해 통계적인 데이터 분석과 시각적인 데이터 분석을 결합하여 시행하였다는 것이다. 통계적인 분석 방법을 통해 예측모형을 도출하였으며, 시각화 분석에서는 다양한 기능을 제공함으로서 최종적으로 제시된 예측모형을 검증하고 데이터로부터 더 다양한 정보를 도출하기 위한 방법론을 제시하였다.

향후 연구로써 본 연구에서 활용한 Parallel Coordinates 방법뿐만 아니라 다양한 시각화 분석 방법을 통계 분석 방법과 결합함으로써 통계적 방법으로 도출하지 못한 데이터의 유의미한 의미를 파악하는 연구가 진행되어야한다.

참고문헌

논문

- David Lechevalier, Anantha Narayanan, Sudarsan Rachuri, "Towards a Domain-Specific Framework for Predictive Analytics in Manufacturing", 2014 IEEE International Conference on Big Data, p. 987, 2014.
- Roiger, R., M. Heatz, "Data mining : A Tutorial Based Primer, Addison Wesley, 2003.
- Soon Tee Teoh, KwanLiu Ma, "Painting Class: Interactive Construction, Visualization and Exploration of Decision Trees", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, p. 668, 2003.
- Adam Perer, Ben Shneiderman,"Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis", CHI 2008 Proceedings · Visual Synthesis, p. 265, 2008.
- E. Kandogan, "Visualizing Multi-Dimensional Clusters, Trends, and Outliers using Star Coordinates.", ACM SIGKDD '01, p. 113, 2001.
- Soon Tee Teoh, KwanLiu Ma, "PaintingClass: Interactive Construction, Visualization and Exploration of Decision Trees", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, p. 670, 2003.
- DeGroot, Schervish, "Definition of a Statistic". Probability and Statistics Third Edition

Addison Wesley, pp.370-371, 2002.
- Pak Chung Wong, J. Thomas, "Visual Analytics", IEEE Computer Graphics and Applications Volume 24 Issue 5, pp. 20, 2004.
- Rick Walker, Philip A. Legg, Serban Pop, Zhao Geng, Robert S. Laramee, Jonathan C. Roberts, "Force-Directed Parallel Coordinates", 17th International Conference on Information Visualisation, p.39, 2013.
- Inselberg, A, The plane with Parallel Coordinates, The Visual Computer, p.79, 1985.
- 문성민, 하효지, 이경원, "영화의 흥행 성과와 리뷰 감정 어휘와의 관계 분석", 디자인 융복합 학회, 제53(4)권, p.7, 2015.
- 경찰청, "지리정보 통합한 지리적 프로파일링 시스템 구축 (GeoPros)", 2013 빅데이터 사례집, p.65, 2013.
- 최종후, 서두성, "의사결정나무를 이용한 개인휴대통신 해지자 분석", 한국경영과학회, pp. 379, 1998.
- 권영란, 김세영, "의사결정나무 분석 기법을 이용한 중학생 인터넷게임중독의 보호요인 예측", 정신간호학회지 13호, p. 19, 2014.

인터넷 사이트

- 영화진흥위원회, http://www.kofic.or.kr/
- http://202.30.24.167:8080/parallel.html