미디어학 석사학위 논문

통계와 시각화를 결합한 데이터분석: 예측 모형 대한 시각화 검증

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

아 주 대 학 교 대 학 원

라 이 프 미 디 어 협 동 과 정

문 성 민

통계와 시각화를 결합한 데이터분석: 예측 모형에 대한 시각화 검증

지도교수 이 경 원

이 논문을 미디어학 석사학위 논문으로 제출함

2015년 12월

아 주 대 학 교 대 학 원

라 이 프 미 디 어 협 동 과 정

문 성 민

문성민의 미디어학 석사학위 논문을 인준함

심사 위원장 이 경 원 인

심 사 위 원 신 현 준 인

심 사 위 원 유 재 인 인

아 주 대 학 교 대 학 원

2015년 12월 15일

요 약

최근 정보통신의 발달과 함께 예측분석의 활용성이 중요해 지고 있으며 이러한 예측분석은 우리의 실생활과 밀접한 관계가 있다. 하지만 예측 분석은 패턴인식(Pattern recognition) 혹은 기계학습(Macine learning)으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 사용자가 분석과정에 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. 또한 본 연구는 통계적인 분석과 시각화 분석을 결합하여 분석을 진행 할 때 영화의 흥행성을 예측하는 것을 목적으로 하였으며 분석을 통해 도출된 결과는 다음과 같다. 첫째, 의사결정나무분석에서 제시된 분할 기준이 적용될 때 마다 변하는 데이터의 패턴을 파악할 수 있다. 둘째, 제시된 최종 예측 모형에 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 영화 흥행성을 예측하기 위한 예측모형을 도출하였으며, 시각화 분석에서는 변수들의 분포를 파악하는 사용자 인터렉션이 가능한 다양한 기능을 제공함으로서 최종적으로 제시된 예측모형을 검증하고 데이터로부터 더 다양한 정보를 도출하기 위한 방법론을 제시하였다.

목 차

그림 목차

표 목차

Ⅰ. 서론

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(machine learning), 데이터마이닝(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 예측한다고 한다. 또한 지리기반 데이터 시각화를 활용하여 예측 분석 결과를 범죄의 유형, 시간대에 따라 범죄다발지역과 위험도를 지도에 각기 다른 색으로 표시하여 수사 과정에 활용한다고 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, 의사 결정 나무 분석 등이 있으며 본 연구에서는 의사결정나무분석을 활용하여 연구를 진행하고자 한다[3]. 의사결정나무분석은 다 변량으로 이루어진 데이터를 분석하기에 적합한 통계적인 예측 분석 방법이며 패턴인식(Pattern recognition) 혹은 기계학습(Macine learning)으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 분석 결과의 정확도와 신뢰성이 높다. 하지만 분석에 사용되는 알고리즘이 복잡하고 많은 조건을 가정해야하는 어려운 분석 일수록 사용자가 분석과정에서 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 이러한 예측

---

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.

분석과 결합하여 분석을 진행함으로써 예측분석의 단점을 보완하고 사용자에게 더 많은 정보를 주기 위한 시도가 이뤄지고 있다. 2008년 발표된 Adam Perer 외 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 의사결정나무분석의 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 제시하고자 한다.

예측분석의 단점을 보완하고 데이터로부터 더 많은 정보를 추출하기 위한 방법론을 제시하기 위해 본 연구에서는 영화에서 추출된 감정어휘 데이터를 활용하여 영화의 흥행성을 예측하는 것을 목적으로 하고자 한다. 영화의 흥행성을 예측하기 위한 기존의 연구로는 1997년 김휴종의 “한국 영화 스타의 스타파워 분석”이라는 연구를 시작으로 김영현 외 1명(2011), 박승현 외 2명(2011)등의 연구가 있다. 해당 연구들은 객관적인 데이터를 활용하여 영화의 흥행성을 예측하기 위한 연구를 진행하였으며 이러한 연구를 통해 감독의 명성, 제작, 유통사의 역량, 평론가의 리뷰, 감독, 작가, 제작자 등 영화의 흥행성에 영향을 미친다는 결과를 제시하였다[7][8][9]. 그러나 객관적인 데이터를 분석한 연구만으로는 영화의 흥행성을 예측하는데 한계점이 존재한다. 선행 연구에서 도출된 영화 흥행성에 영향을 미치는 요인들의 조건을 충족하지만 영화 흥행에 실패한 영화의 예로 2014년 개봉한 영화 ‘익스펜더블3’가 있다. 이 영화의 경우 유명한 해외 스타들이 다수 등장하고 유명한 감독이 연출하고 해외에서 이미 흥행에 성공을 했지만 국내 시장에선 흥행(누적 관객 152,025명)에 성공 하지 못했다. 이러한 사례처럼 객관적인 데이터만으로는 영화의 흥행성을 예측하는데 한계가 존재하며 객관적인 데이터를 활용한 분석이 아닌 다른 관점에서의 분석이 필요하다. 객관적인 데이터의 한계를 보완하기 위해 본 연구에서는 관객들이 영화를 통해 느낀 감정과 같이 영화리뷰에서 추출 할 수 있는 주관적인 데이터를 활용하여 영화 흥행성을 높이기 위한 분석을 수행하고자 한다.

본 연구는 통계적인 분석방법으로 영화의 흥행성을 예측하기 위해서 의사결정

---

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.

7 김휴종, “한국영화스타의 스타파워분석”, 삼성경제연구소 연구보고서, 1997.

8 김연형, 홍정한, “영화 흥행 결정 요인과 흥행 성과 예측 연구”, 한국통계학회논문집, 제18권, 제6호, p.865, 2011.

9 박승현, 송현주, 정완규, “한국영화의 흥행성과 결정 요인에 관한 연구”, 언론과학연구, 제11권, 제4호, p.240, 2011.

나무분석을 사용할 것이고 최종 제시된 예측모형에 대한 검증하기 위해 시각화 분석을 사용하고 두 데이터 분석 방법을 결합하여 사용하는 방법을 제시하고자 한다.

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 제시하고자 한다. 이를 위해 영화 리뷰에서 추출되는 감정 어휘 데이터를 기반으로 영화 흥행성을 예측을 위한 의사결정 나무분석을 수행하고 시각화 분석방법을 활용하여 분석된 예측모형을 검증하고 데이터로부터 더욱 다양한 정보를 도출하는 방법을 제시하고자 한다.
연구 목적을 달성하기 위한 연구 진행 과정은 다음과 같다.
첫째, 네이버 영화 평에서 리뷰의 개수가 1000개 이상인 672개의 영화에 대한 리뷰를 크롤링(Crawling)하고 감정어휘 사전을 활용하여 감정어휘 데이터를 생성하였다.
둘째, 감정어휘 데이터를 기반으로 비슷한 감정이 느껴지는 영화 장르를 군집화 하였다. 또한 군집된 결과를 MDS 시각화를 활용하여 검증하였다.
셋째, 높은 영화 흥행도 예측 값을 도출하기 위해 전체 영화와 군집화를 통해 생성된 4개의 복합장르 그룹에 따라 의사 결정 나무 분석을 시행하였다.
넷째, 다양한 시각에서 데이터를 분석하기 위해 Parallel coordinate 시각화를 제작하고 시각화 분석 방법을 활용하여 데이터를 분석 하였다.
다섯째, Parallel coordinate 시각화를 활용하여 의사결정나무분석에서 제시된 최종 모형에 대한 검증을 수행하였다.
여섯째, 분석 결과를 해석하고 연구의 시사점과 연구의 한계, 향후 연구 방향을 제시하였다.

Ⅱ. 이론 및 선행 연구의 고찰

1. 영화 흥행 관련연구

영화 산업 시장의 규모가 확대됨에 따라 영화의 흥행에 영향을 미치는 요인을 도출해 내기 위한 다양한 연구들이 진행 되어오고 있다.

국내에서는 1997년 김휴종의 '한국 영화 스타의 스타파워 분석'이라는 연구와 2011년 박승현, 송현주, 정완규의 '한국영화의 흥행성과 결정요인에 관한 연구' 등 다양한 연구가 있다. 김휴종의 연구에서는 1988년부터 1995년까지의 529편의 영화를 회귀분석을 사용하여 분석하였으며 배우와 감독의 스타 파워가 흥행에 미치는 영향력을 검증하였다[10]. 다음으로 박승현 외 2명의 연구에서는 회귀분석을 사용하여 개봉 스크린 규모, 제작비, 전문가 평가, 온라인 평가 등이 영화 흥행에 영향을 미치는 것을 검증하였다[11].

해외에서는 1983년 리트만(Litman)의 'Predicting Success of Theatrical Movies: An Empirical Study'와 1991년 와이어트(Wyatt)의 'High concept, product differentiation, and the contemporary U.S film industry'등의 연구가 있다. 리트만의 연구는 1970년대 개봉한 영화 155편에 대한 정보를 수집하고 회귀분석을 실시하여 제작비, 개봉스크린 규모, 전문가 평점, 배급사 파워, 아카데미 수상실적 등의 요소와 공상 과학/판타지, 코미디, 공포의 세 가지 장르 요소가 유의미한 영향력을 지닌다는 결과를 도출하였고[12] 와이어트의 연구는 제작비, 전문가 평점, 스타 배우의 파워, 아카데미 수상실적, 여름시즌 개봉의 요인이 영화 흥행에 영향을 미친다는 결과를 도출하였다[13]. 영화 흥행도 관련 선행 연구의 내용은 표 1 과 같다.

저자	논문	흥행에 영향을 미치는 요인
김휴종	한국영화스타의 스타 파워 분석(1997)	배우의 스타 파워, 감독의 스타 파워
박승현 외 2명	한국 영화의 흥행성과 결정 요인에 관한	개봉 스크린 규모, 제작비, 전문가 평가, 온라인 평가

10 김휴종.

11 박승현, 송현주, 정완규.

12 Litman, B, "Predicting Success of Theatrical Movies: An Empirical Study", Journal of Popular Culture, 16 (Spring), p.166, 1983.

13 Wyatt, R. O, "High concept, product differentiation, and the contemporary U.S film industry", Current research in film Audiences, economics, and law, Vol. 5, p.93, 1991.

	연구(2011)	
송현주 외 1명	영화의 흥행성과와 제작비 규모와의 관계(2012)	온라인 평가, 영화 평점, 제작비, 코미디 장르
이윤정 외 1명	원작의 유무와 형태가 영화 흥행에 미치는 영향(2013)	소설 원작의 존재, 소설 원작의 형태, 소설 원작의 영향력, 영화 관련 외부 콘텐츠
Litman, B	Predicting Success of Theatrical Movies: An Empirical Study(1983)	제작비, 개봉스크린 규모, 전문가 평점, 배급사 파워, 아카데미 수상실적, 공상 과학/판타지, 코미디, 공포(3가지 장르)
Wyatt, R. O	High concept, product differentiation, and the contemporary U.S film industry(1991)	제작비, 전문가 평점, 스타 배우의 파원, 아카데미 수상실적, 여름시즌 개봉

표 1. 영화 흥행도 관련 논문

2. 영화 감정 어휘 관련연구

주관적인 데이터는 영화를 관람한 관객에 의해 생성되며 영화에 대한 평가 리뷰 안에서 감정 어휘를 추출하여 분석하면 영화 흥행과 감정 어휘 사이의 관계를 파악 할 수 있다. 감정 어휘에 대한 연구는 언어에 따라 연구 방법이 상이한데 한글의 텍스트 기반 감정 언어에 대한 연구는 2008년 이준웅, 송현주, 나은경, 김현석의 연구가 있고 영화 리뷰를 활용하여 감정 어휘를 분석한 연구는 2014년 박지연, 전범수의 연구가 있다.

이준웅 외 3명의 연구(2008)에서는 유사성 분류 자료를 근거로 군집 분석을 수행하였고, 감정 어휘를 기본 수준에서 '기쁨', '긍지', '사랑' 등 긍정적인 정서들과 '공포', '분노', '연민', '수치', '좌절', '슬픔' 등 부정적인 정서로 총 9개의 정서 범주로 나눌 수 있음을 제시하였다. 또한 '기쁨'의 경우 다시 '기쁨'과 '통쾌'로 나누어지며 나누어진 '기쁨'도 '재미', '즐겁', '유쾌', '흥겹', '신나', '기쁘', '열망'으로 나누는 등 대표 감정 어휘가 포함하는 세부 감정 어휘에 대한 온톨로지를 구축하였다[14].

박지연 외 1명의 연구(2014)에서는 한국 및 외국 흥행 영화에 대한 네티즌 리뷰를 중심으로 네티즌 리뷰에 사용된 감정 동사와 흥행 영화와의 관계를 분석

14 이준웅, 송현주, 나은경, 김현석, "정서 단어 분류를 통한 정서의 구성 차원 및 위계적 범주에 관한 연구", 한국 언론 학보, 제52(1)권, p.101, 2008.

하였다. 연구 결과로는 네티즌들이 영화를 판단 할 때 재미를 가장 큰 요인으로 생각하고 있다는 점과 한국 영화의 경우 재미를 기준으로 몰입이나 감동 등 감정적 동사를 기준으로 영화가 군집화 되고 외국 영화는 재미있는 영화와 재미없는 영화로 군집화 된다는 결과를 얻었다[15].

3. 의사결정나무분석 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 의사결정나무분석은 다 변량으로 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 '의사결정나무분석 기법을 이용한 중학생 인터넷게임중독의 보호요인 예측' 등의 연구가 있다.

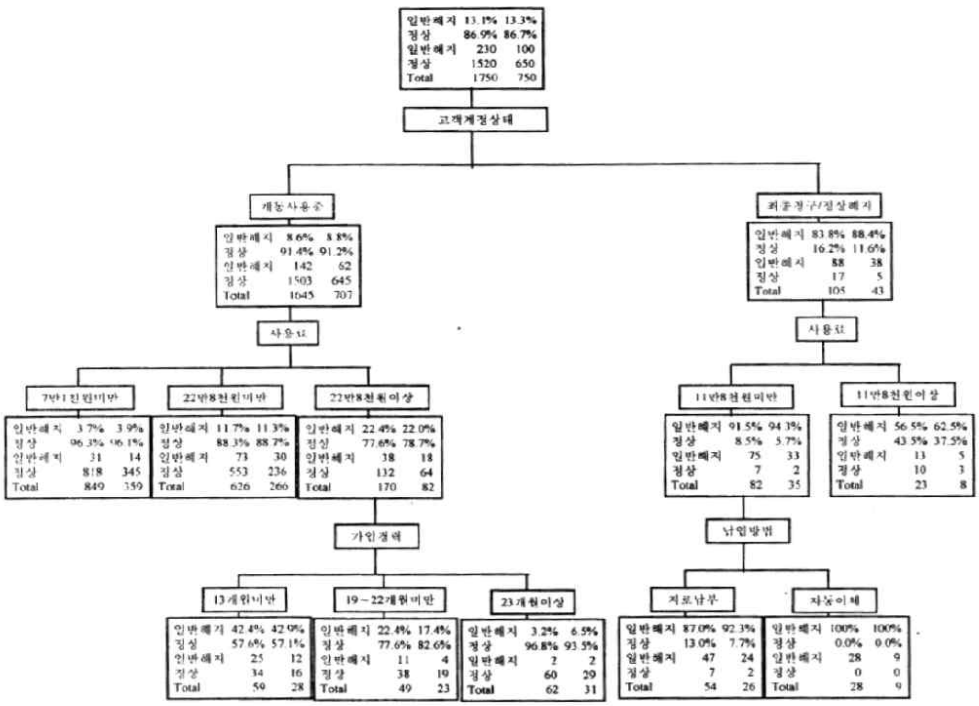

그림 1. 최종후 외 1명 연구의 <그림1> "의사결정나무"

15 박지연, 전범수, "네티즌의 흥행 영화 리뷰에 포함된 감정 동사 이용 특성 연구", 한국콘텐츠 학회, 제14(5)권, p.88, 2014.

최종후, 서두성의 연구에서는 휴대전화 가입 고객의 해지를 결정하는 제일 중요한 변수는 고객계정상태이며, 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[16].

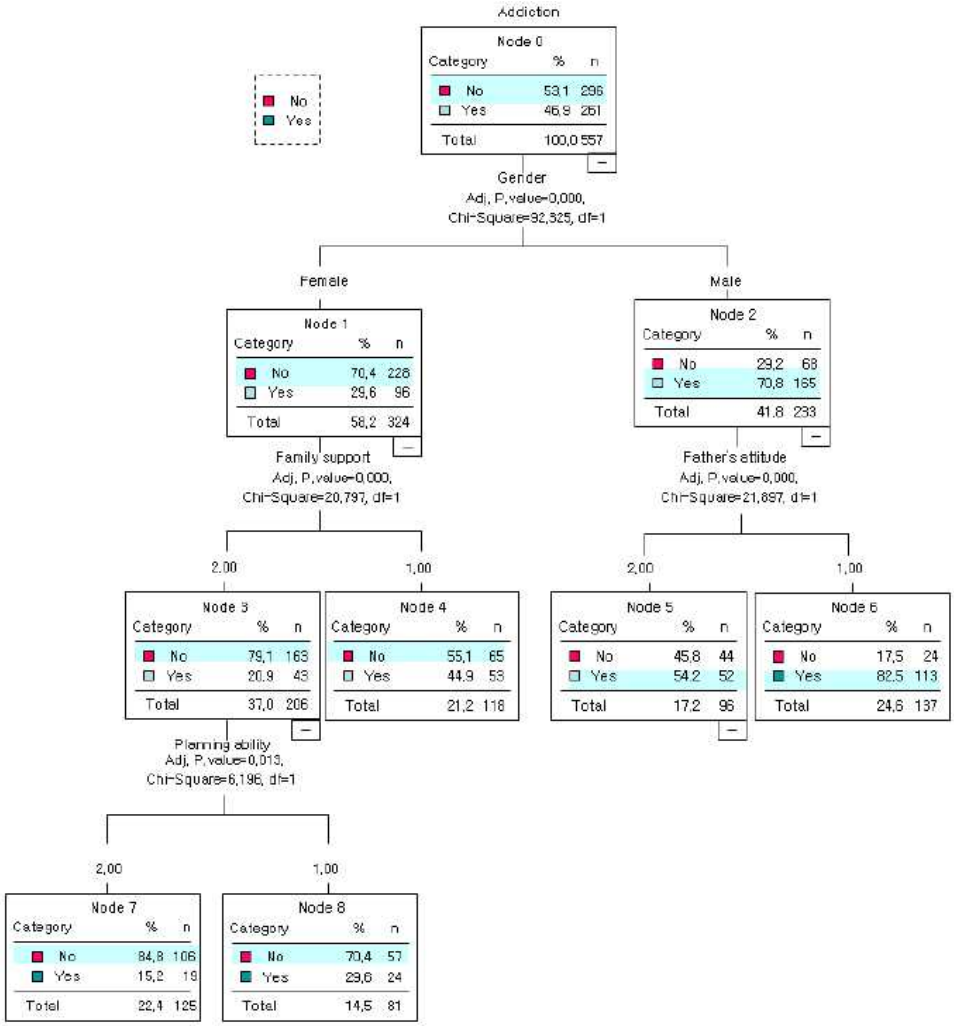

그림 2. 권영란 외 1명 연구의 <그림1> “The construction of decision tree.”

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

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

남학생의 경우에는 아버지의 태도가 엄격할수록 인터넷게임 일반 사용군에 포함될 확률이 높다는 결과를 제시하였다[17].

4. 시각화분석 관련연구

의사결정나무분석은 다 변량으로 이루어진 데이터를 분석하기에 적합한 통계적인 분석 방법이다. 하지만 패턴인식(Pattern recognition) 혹은 기계학습(Macine learning)으로 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 'PaintingClass: Interactive Construction, Visualization and Exploration of Decision Trees'이라는 연구가 있다.

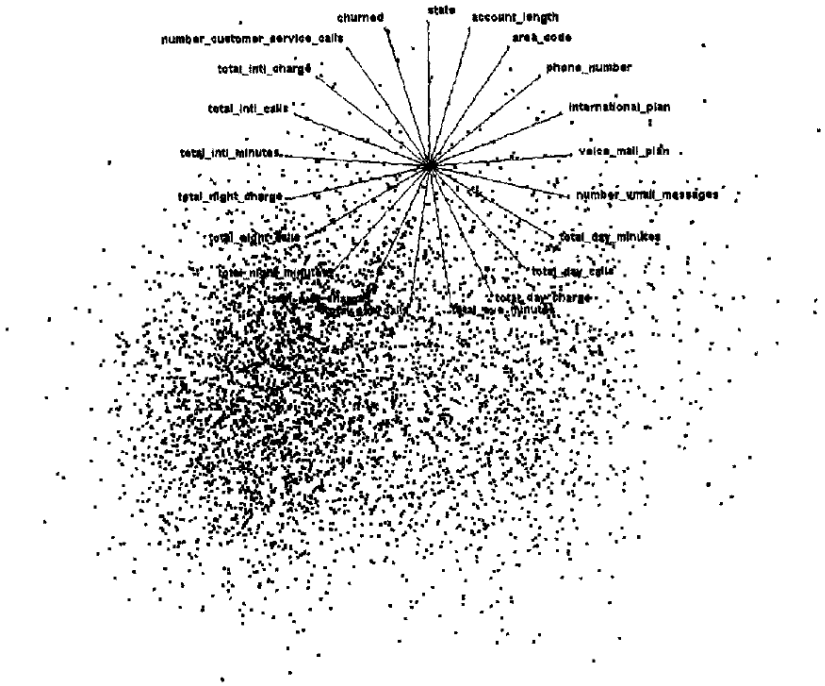

그림 3. Eser Kandogan 연구의 <그림12> "Overview of 'churn' dataset, where churned customers are marked with blue (dark) color."

Eser Kandogan의 연구에서는 시각화 분석 방법 중 하나인 Star Coordinates

17 권영란, 김세영, "의사결정나무분석 기법을 이용한 중학생 인터넷게임중독의 보호요인 예측", 정신간호학회지 13호, p. 19, 2014.

활용하여 다 변량의 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[18].

Soon Tee Teoh외 1명의 연구에서는 의사결정나무분석 결과를 Parellel coordinates와 Star Coordinates와 같은 시각화 분석 방법을 통해 나타냄으로써 데이터에서 발견할 수 있는 결과를 폭 넓게 도출하고자 하였다. 또한 이 두 시각화를 연결하여 사용하면 데이터 분류 과정과 데이터의 분류를 통합하여 확인 할 수 있다는 제안을 하였다[19].

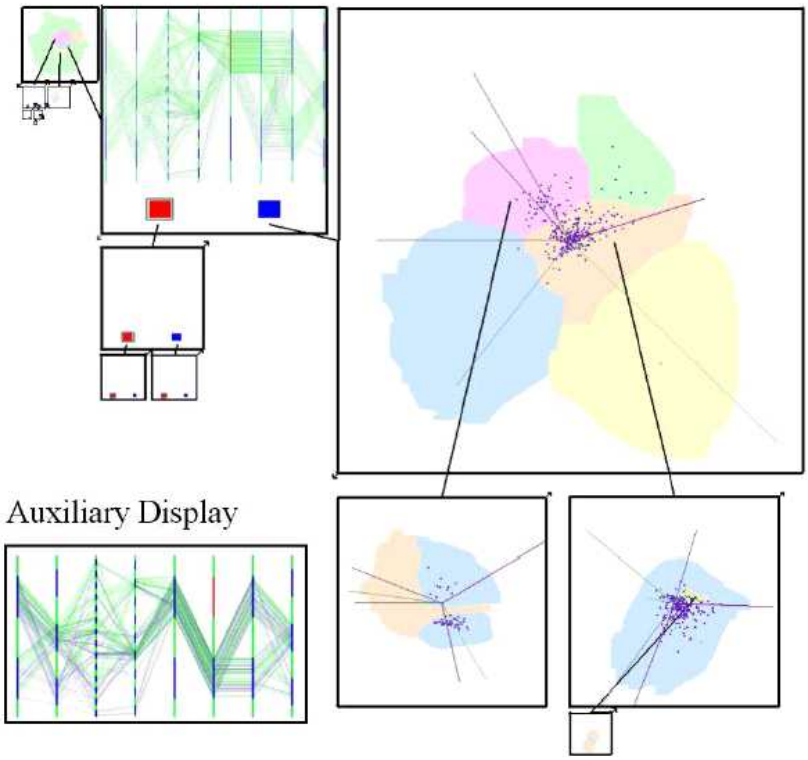

그림 4. Soon Tee Teoh 외 1명 연구의 <그림5> “An auxiliary display is shown on the un-utilized space at the lower left of the display.”

---

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

19 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.

Ⅲ. 데이터 수집 및 정제

1. 영화 리뷰 데이터 수집

본 연구의 수행을 위해 아래와 같이 두 단계의 데이터 수집 과정을 거쳤다.
첫째, 자동화된 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.

2. 감정어휘 사전 구축 및 감정어휘 데이터 생성

본 연구는 선행 연구 중 하효지 외 2명의 연구(2013)를 참고하여 감정 어휘 사전을 구축하기 위해 여러 방법을 거쳤으며 감정 어휘 사전 구축 과정은 다음과 같다[20].
첫째, 한덕웅, 강혜자(2000)의 한국어 정서 용어들의 적절성과 경험 빈도에 관한 연구를 참고하여 834개의 정서용어 중에서 영화를 봤을 때 느낄 수 있는 감정어휘만을 분류하는 작업을 시행하였다[21]. 작업은 국어국문학 분야의 전문가의 견해를 바탕으로 진행되었으며 이를 통해 최종 100개의 감정어휘를 선별하였다. 또한 최종 감정 어휘 사전을 구축하기 위한 설문 조사를 실시하여 68개의 감정어휘를 최종 선정하게 되었다. 최종 선정된 68개의 감정어휘는 표 2 와 같다.

68개의 감정어휘					
가련하다	가슴아프다	감격스럽다	감동적이다	거북하다	겁나다
격분하다	경악하다	경이롭다	경쾌하다	공포스럽다	굉장하다
구역질나다	그립다	기겁하다	기쁘다	끔찍하다	나른하다
놀라다	늘어지다	달콤하다	담담하다	대단하다	더럽다
두렵다	등골이 서늘하다	무료하다	무섭다	무시무시하다	분노하다
분하다	불결하다	불쌍하다	불쾌하다	불편하다	서글프다

20 하효지, 김기남, 이경원, “영화 리뷰의 감정 어휘 공간 및 영화 관람의 상황분석 연구”, 디자인 융복합 학회, 제12(6)권, p.51, 2013.
21 한덕웅, 강혜자, “한국어 정서 용어들의 적절성과 경험 빈도.” 한국 심리학회지: 일반, Vol.19, 2000, pp.90.

서럽다	섬뜩하다	소름끼치다	속상하다	슬프다	신경질나다
신나다	쓸쓸하다	안타깝다	암울하다	애석하다	애절하다
역겹다	오싹하다	우울하다	유쾌하다	으스스하다	잔인하다
재미있다	즐겁다	증오스럽다	지루하다	징그럽다	차분하다
통쾌하다	평온하다	행복하다	혐오스럽다	환상적이다	활기 있다
황홀하다	흥겹다				

표 2. 하효지 외 2명의 연구, “최종 68개의 감정어휘”

둘째, W. Parrot(2001)의 Emotion Table[22]에서 제시한 6가지 기본 감정에서 ‘Boring(지루한)’을 더하여 영화에 적합한 7가지의 대표 감정어휘를 생성하고 대표 감정어휘에 의해 68개의 감정어휘를 군집화 하였다. 군집화 된 감정어휘는 아래 표 3 과 같다.

68개의 감정어휘						
Happy	Sad	Anger	Surprise	Disgust	Fear	Boring
경쾌하다	가련하다	격분하다	감격스럽다	거북하다	겁나다	평온하다
기쁘다	가슴 아프다	분노하다	감동적이다	구역질나다	공포스럽다	나른하다
달콤하다	그립다	분하다	경악하다	끔찍하다	두렵다	차분하다
신나다	불쌍하다	불편하다	경이롭다	더럽다	등골이 서늘하다	담담하다
유쾌하다	서글프다	신경질나다	굉장하다	불결하다	무섭다	늘어지다
재미있다	서럽다	증오스럽다	기겁하다	불쾌하다	무시무시하다	지루하다
즐겁다	속상하다		놀라다	역겹다	섬뜩하다	무료하다
통쾌하다	슬프다		대단하다	잔인하다	소름끼치다	
행복하다	쓸쓸하다		환상적이다	징그럽다	오싹하다	
활기있다	안타깝다			혐오스럽다	으시시하다	
황홀하다	암울하다					
흥겹다	애석하다					

22 Parrott, W. Emotions in Social Psychology. Philadelphia: Psychology Press, 2001.

	애절하다					
	우울하다					

표 3. 하효지 외 2명의 연구, “감정어휘의 군집화”

셋째, TF-IDF공식을 활용하여 감정 어휘 데이터를 표준화시켰다. TF-IDF는 각 감정어 집단의 단어 빈도수(tf : Term Frequency)와 역 문서빈도(idf : Inverse Document Frequency)를 곱하여 감정 어휘를 표준화시킨 공식으로써 TF-IDF값에 관한 공식은 다음과 같다.

$$TF-IDF(t,d,D) = tf(t,d)*idf(t,D)$$

감정 어휘 개수를 줄이기 위해 각 감정 어휘에서 나타날 수 있는 TF-IDF 스코어의 최대치를 구하였다. 예를 들어 ‘경악하다’의 경우 모든 영화에서 TF-IDF 스코어의 비율이 0.8% 이하인 반면에 ‘달콤하다’의 경우는 적어도 한 개의 영화에서는 TF-IDF 스코어의 비율이 42%에 달하는 것을 뜻한다. 따라서 TF-IDF스코어의 비율이 10%미만인 감정 어휘를 제거하고 최종적으로 36개의 감정 어휘를 선택하였다. 최종적으로 선택된 36개의 감정 어휘는 크게 ‘Happy’, ‘Surprise’, ‘Boring’, ‘Sad’, ‘Anger’, ‘Disgust’, ‘Fear’의 성격으로 나뉘게 되며, 감정 어휘에 관한 내용은 표 4 와 같다.

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)
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)

표 4. 하효지 외 2명의 연구, “최종 선정된 36개의 감정 어휘”

본 연구는 주관적인 데이터로써 영화를 보고 관객들이 느끼는 감정을 기반으로 영화의 흥행성을 예측하고자 하였으며 이를 위해 앞의 과정을 통해 얻은 데이터 중 대표 감정 어휘와 각 영화의 객관적인 데이터를 병합하여 최종적인 데이터를 생성하였다. 병합된 영화에 대한 객관적인 데이터는 영화 티켓 판매액, 영화 관람 관객 수, 상영 스크린 수, 한 스크린 당 영화 관람 관객 수, 영화의 장르, 영화의 영문 이름 등이었으며 영화진흥위원회의 통계 결과를 바탕으로 데이터를 생성하였다[23]. 영화 흥행도를 나타내는 요인으로 1983년 리트만(Litman)의 연구에서는 누적 매출액을 영화 흥행도로 사용하였으며,[24] 최근 연구들 중 2012년 박승현 외 1명의 연구에서는 누적 관객 수를 영화 흥행도를 나타내는 요인으로 사용하였다[25]. 본 연구에서는 누적 관객 수를 상영 스크린 수로 나누어 한 스크린에서의 누적 관객 수를 영화 흥행도로 사용하였으며 위의 데이터와 병합하여 분석에 사용하였다.

3. 데이터 특성 파악

672개의 영화에 대한 총 판매액의 평균은 10,190,000,000원이고 누적 관객 수의 평균은 1,469,038명, 평균 상영 스크린은 330.7개 상영관, 흥행도를 나타내는 한 스크린 당 누적 관객 수의 평균은 4,018명이었다. 마지막으로 672개의 영화에 대한 영화 감정어휘의 평균은 ‘Happy’(0.3073), ‘Surprise’(0.2625), ‘Sad’(0.1465), ‘Boring’(0.09126), ‘Fear’(0.08708), ‘Anger’(0.06247), ‘Disgust’(0.04314)순으로 ‘Happy’와 ‘Surprise’의 평균 감정어휘 값이 다른 감정 어휘의 평균값보다 높았다. 최종적으로 분석에 사용 될 데이터의 특징은 표 5 의 내용과 같다.

변수명	영문명	최대값	평균	최솟값
판매액	Sales	1.280e+11	1.019e+10	1.670e+04
관객수	Attendance	13624328	1469038	35
개봉스크린수	Screen	1409.0	330.7	1.0
흥행도(평균 관객수)	Nomal_Attendance	33590	4018	35

23 영화진흥위원회.

24 Litman, B.

25 박승현, 송현주, “영화의 흥행성과와 제작비 규모와의 관계: 2011년 한국영화의 흥행결정요인 분석”, 사회과학연구, 제51집 1호, p.67, 2012.

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

표 5. 영화 데이터의 특징

Ⅳ. 통계를 활용한 예측 분석

최근 영화의 흥행성에 영향을 미치는 요인을 도출하고자 선행된 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 α혹은 p(probability)<.05의 수준으로 정하고 통계분석 결과가 이를 만족하면 유의미한 결과라고 해석한다[26]. 하지만 분석에 사용되는 알고리즘이 복잡하고 많은 조건을 가정해야하는 어려운 분석 일수록 사용자가 분석과정에 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[27]. 최근에는 이러한 단점을 보완하고 데이터로부터 더 많은 정보를 얻어내기 위해 시각화 분석을 결합하여 분석을 진행하고 있다[28]. 시각화 분석이란 데이터 분석 결과를 쉽게 이해할 수 있도록 시각적으로 표현하고 전달하는 분석 방법으로써 연결과 그룹화를 통한 데이터 요약, 색, 모양 등 미적 요소를 활용한 데이터의 특성 표현 등 다양한 방법으로 사용자의 이해를 돕는다[29]. 본 연구는 통계적 분석과 시각화 분석을 결합하여 연구를 진행하였다. 이를 위해 통계적 분석을 수행 후 분석 결과에 대해 시각화 분석을 수행함으로써 이를 검증하였다.

1. 군집분석과 MDS시각화를 이용한 영화 장르 군집화

영화 시장의 성장과 함께 다양한 장르의 영화들이 제작되었으며 최근에는 복합장르의 영화도 많이 제작되고 있다. 본 연구는 각 영화에 대해 대표 장르 별로 영화를 구분하여 데이터를 생성하였으며 최종 선정된 영화의 장르는 “Drama, Action, Comedy, Meloromance, SF, Horror, Thriller, Crime, Fantasy, Mystery, Historicaldrama, Family, Adventure, War, Sports”로 15개의 장르가 선정되었다. 하지만 시각화 분석 결과 장르가 상이하여도 영화를 통해 관객이 느끼는 감정값의 분포가 비슷한 장르들이 있다는 것을 확인하였다. 따라서 본 장에서는 감정 어휘 값을 기반으로 15가지 장르에 대해 통계적인 분석 방법 중 군집분석 방법을 수행하여 영화의 장르를 군집화 하고자 한다. 또한 군집된 집단에 대해 MDS를 활용한 시각화 분석을 수행하여 군집간의 유사성을

---

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

[27] Soon Tee Teoh.

[28] Adam Perer.

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

확인하고자 한다.

가. 군집분석의 정의

군집 분석(cluster analysis)이란 N개의 관찰치를 대상으로 p개의 변수를 측정했을 때, 관측한 p개의 변수 값을 이용하여 N개의 관찰치 사이의 거리(distance)를 측정하여 관찰치(N)들을 군집화 하는 통계적 분석 방법이다[30]. 본 장에서는 각 장르별 영화의 감정어휘 값에 대해 유클리디안 거리 공식을 활용하여 유사도 값을 측정하였다. 사용된 유사도 측정 공식은 다음과 같다.

$$\text{Euclidean distance} = d(D_i, D_j) = \sqrt{\sum_{i=1}^{n}(D_i - D_j)^2}$$

감정어휘 값을 기반으로 장르별 유사도를 측정한 후에는 최장거리(furthest-neighbor) 연결법과 최단거리 연결법(nearest-neighbor)을 사용하여 군집화 하였다. 분석에 사용된 군집 연결법 공식은 다음과 같다.

$$\text{Nearest-neighbor function} = d_{\min}(D_{i,}D_j) = \min \parallel X - X' \parallel$$

$$\text{Furthest-neighbor function} = d_{\max}(D_{i,}D_j) = \max \parallel X - X' \parallel$$

나. 군집분석 결과

본 장에서는 15개의 영화 장르들의 평균적인 감정어휘 값을 계산하고 이를 기반으로 감정어휘의 분포가 비슷한 영화의 장르끼리 군집화 하고자 하였다. 분석에는 장르별 영화의 감정어휘 유사도 값을 구하는 공식으로 유클리디안 거리 공식을 사용하여 측정하고 최장거리(furthest-neighbor) 연결법과 최단거리 연결법(nearest-neighbor)을 사용하여 군집을 연결하였다. 분석 결과는 그림 5 , 그림 6 과 같다.

[30] Wikipedia, Cluster analysis, https://en.wikipedia.org/wiki/Cluster_analysis

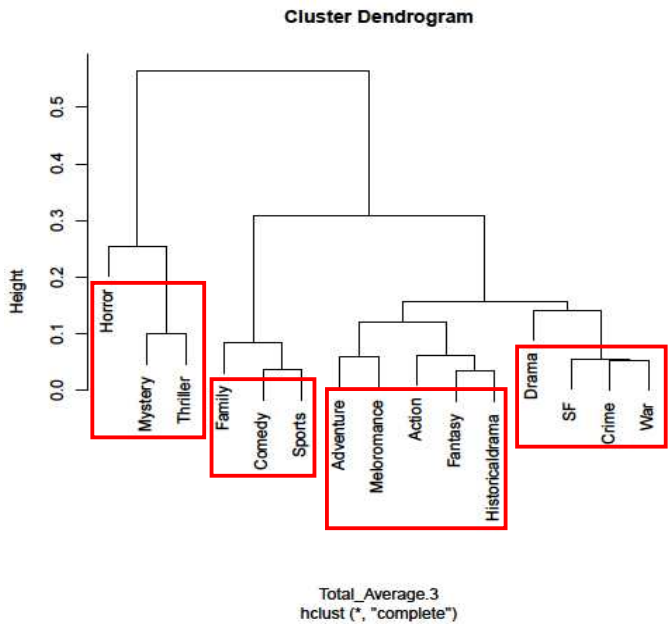

그림 5. 유클리디안 거리, 최장거리 연결법 수행을 수행한 수형도

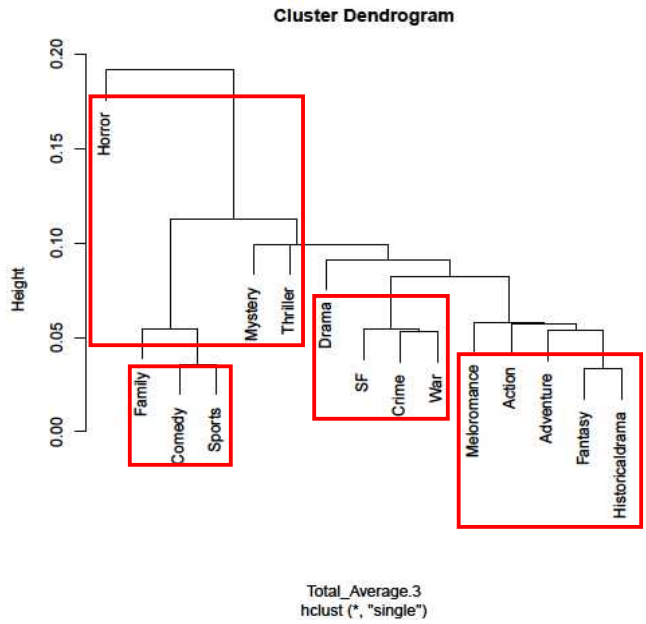

그림 6. 유클리디안 거리, 최단거리 연결법 수행을 수행한 수형도

군집 분석을 수행 수 수형도(Dendrogram)를 통해 분석 결과를 확인한 결과 총 4개의 군집으로 영화의 장르가 군집화 되는 것을 확인하였다. 1번 군집의

경우 Horror, Mystery, Thriller가 속하였으며 2번 군집에는 Family, Comedy, Sports, 3번 군집에는 Adventure, Meloromance, Action, Fantasy, Historicaldrama, 그리고 4번 군집에는 Drama, SF, Crime, War이 속하였다.

다. 군집 결과 시각화

다차원척도법 (Multi-Dimensional Scaling)으로 많이 알려진 MDS시각화는 데이터 사이의 관계에 관한 수치적 자료를 처리하여 다차원 공간상에서 그 대상들을 위치적으로 표시하여 주는 분석 방법으로써 전체적인 관계구조를 공간상의 그림을 통해 쉽게 파악할 수 있게 한다[31]. 군집 분석 결과를 MDS시각화를 활용하여 나타낸 결과는 그림 7 과 같다.

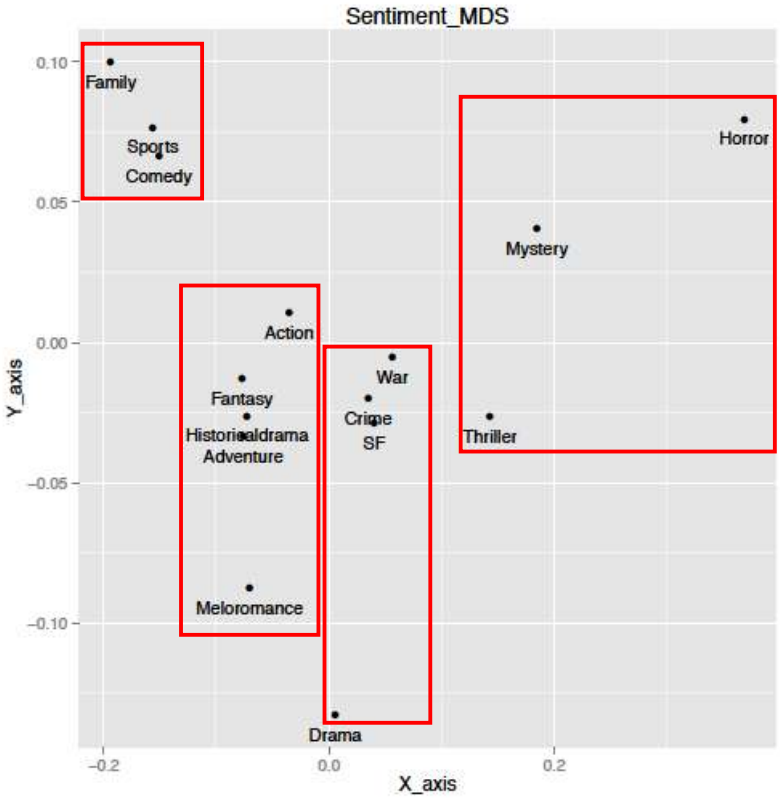

그림 7. MDS시각화 분석 결과

MDS시각화 분석 방법을 활용한 결과, 총 4개의 군집으로 영화의 장르가 군집화 되는 것을 확인하였으며 노드간의 거리를 통해 군집 내 장르 중에서도 더 유사한 장르가 어떠한 장르인지 확인 할 수 있다.

[31]Borg, I., Groenen, P., "Modern Multidimensional Scaling: theory and applications", New York: Springer-Verlag, pp. 208, 2005.

2. 의사결정나무분석을 활용한 영화 흥행도 예측

가. 의사결정 나무 분석의 정의

의사결정나무는 의사결정규칙(decision rule)을 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(classification)하거나 예측(prediction)을 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[32]. 의사결정나무는 분류 또는 예측을 목적으로 하는 어떤 경우에도 사용 될 수 있으나 분석의 정확도보다는 분석과정의 설명이 필요한 경우에 더 유용하게 사용된다. 의사결정나무 분석이 활용될 수 있는 응용분야는 표 6 과 같다.

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

표 6. 의사결정나무분석 응용분야

[32] 손용정, 김현덕.

의사 결정 나무분석은 목표변수, 예측변수, 분리기준, 분리개수에 따라 크게 CHAID(Chi-squared Automatic Interaction Detection), Exhaustive CHAID, CART(Classification and Regression Trees), QUEST(Quick Unbiased Efficient Statistical Tree)로 나누어진다. 언급된 네 가지 분석 방법에 대한 설명은 표 7 과 같다.

	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

표 7. 의사결정나무분석의 종류

본 연구에서 분석에 사용될 영화 흥행도와 7가지 대표 감정어휘 값의 경우 연속 형으로 이루어진 데이터 세트이다. 따라서 네 가지의 분석 방법 중 목표변수(종속변수)와 예측변수(독립변수)로 연속 형 데이터(양적변수)를 다루고, 분리개수가 이지분리를 따르는 CART(Classification and Regression Trees) 분석 방법을 사용하였다. CART(Classification and Regression Trees)는 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(연속 형 목표변수인 경우 적용)을 이용하여 이지분리(binary split)를 수행하는 알고리즘이다. 가장 널리 사용되는 의사결정나무 알고리즘으로 개별 입력변수 뿐만 아니라 입력변수들의 선형 결합들 중에서 최적의 분리를 찾을 수도 있다.

나. 전체 영화에 대한 의사결정나무분석

전체 영화 데이터 집단에 대한 의사 결정 나무분석을 수행하기에 앞서 목표가 되는 종속 변수와 분할기준으로 작용을 할 독립변수에 대한 기술 통계 값은 표 8 과 같다.

변수구분	변수명	평균	최대값	최솟값
종속변수	흥행도 (Nomal-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

표 8. 분석에 사용된 데이터의 특징(전체영화)

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

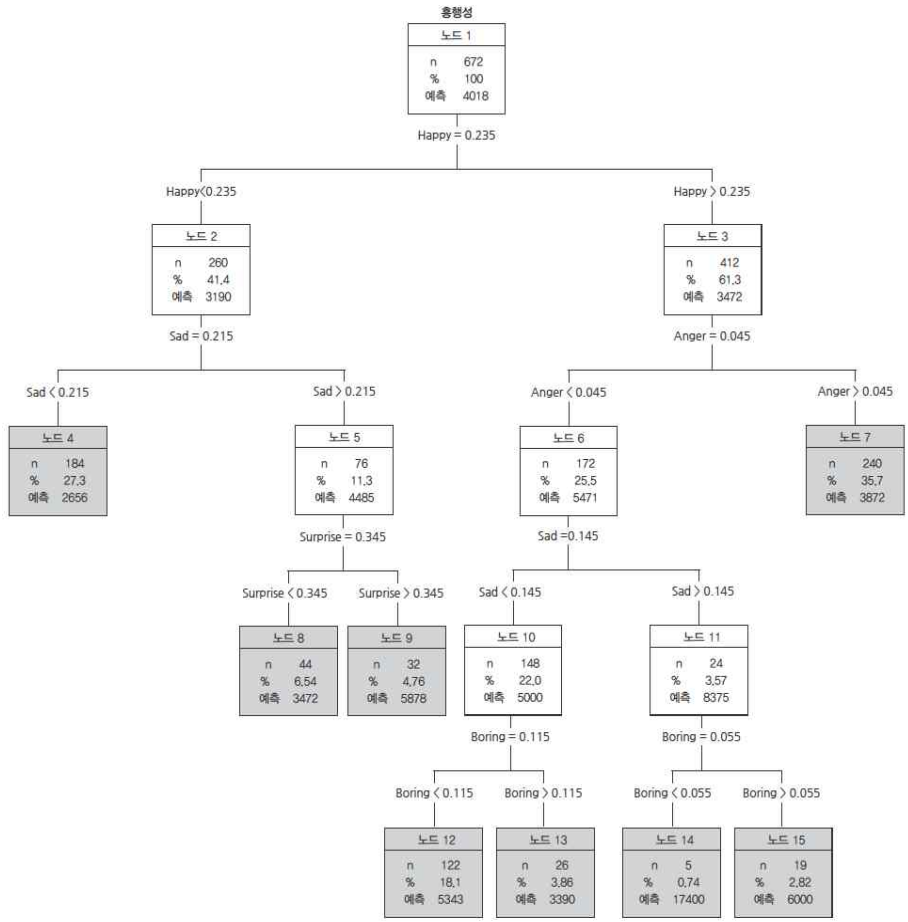

그림 8. 전체 영화에 대한 의사 결정 나무 분석

전체 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 영화 흥행성이 가장 높다고 예측된 집단에 대한 해석은 다음과 같다. 영화 흥행성이 17400이 되기 위해서는 Happy 0.235를 기준으로 최조 분리되어야 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 흥행성의 예측 값은 17400으로 높게 분류된다. 분석된 의사결정 나무분석 결과는 그림 8 과 같다.
의사결정 나무 분석의 결과는 자료의 분류가 얼마나 잘되었는지 한눈에 표현하는 이익도표를 통해 더 자세히 확인 할 수 있다. 전체 영화 집단에 대한 이익도표 값은 표 9 와 같다.

노드번호	개수(N)	비율(%)	영화 흥행성 예측값
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

표 9. 전체 영화 집단에 대한 이익도표

다. 군집 1 영화에 대한 의사결정나무분석

군집 1 집단에 대한 의사 결정 나무분석을 수행하기에 앞서 목표가 되는 종속변수와 분할기준으로 작용을 할 독립변수에 대한 기술 통계 값은 표 10 과 같다.

변수구분	변수 명	평균	최댓값	최솟값
종속변수	흥행도 (Nomal-Attendance)	2904.3	12908.26	87.69
독립변수	Happy	0.1737	0.49	0.04
	Surprise	0.2146	0.59	0.06
	Boring	0.0764	0.32	0.02
	Sad	0.1169	0.27	0.05
	Anger	0.0632	0.28	0.01
	Disgust	0.0675	0.37	0.00
	Fear	0.2882	0.65	0.01

표 10. 분석에 사용된 데이터의 특징(군집1)

분석에 사용된 군집1 집단에 대한 기술 통계량 값을 보면 흥행도는 종속 변수로써 평균값이 2904.3이며 7개의 대표 감정 어휘 값들이 독립변수로 사용되었고 Fear와 Surprise, Happy의 감정어의 평균값이 높았으며 그 중에서도 Fear 값이 가장 높은 것을 확인 할 수 있다.

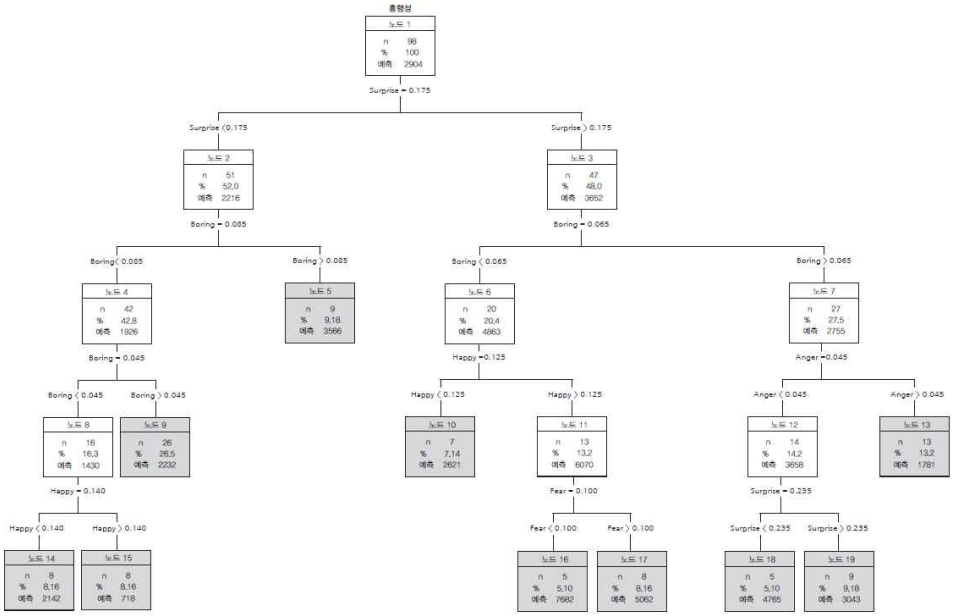

그림 9. 군집 1에 속한 영화에 대한 의사 결정 나무 분석

군집1 에 속한 영화 데이터 집단에 대한 최적분리는 Surprise에 의해 최조 이지 분리 되었다. 영화 흥행성이 가장 높다고 예측된 집단에 대한 해석은 다음과 같다. 영화 흥행성이 7682가 되기 위해서는 Surprise 0.175를 기준으로 최조 분리되어야 하며 Surprise가 0.175이상 일 경우 Boring 0.065를 기준으로 다시 분리된다. Boring이 0.065이하 일 경우 다시 Happy 0.125를 기준으로 분리되며 Happy가 0.125 이상 일 경우 마지막으로 Fear 0.100을 기준으로 분리되며 Fear가 0.100이하 일 때의 집단(N=5)에 대한 영화 흥행성의 예측 값은 7682로 높게 분류된다. 분석된 의사결정 나무분석 결과는 그림 9 와 같다.
의사결정 나무 분석의 결과는 이익도표를 통해 더 자세히 확인 할 수 있다. 군집 1에 속한 영화 집단에 대한 이익 도표 값은 표 11 과 같다.

노드번호	개수(N)	비율(%)	영화 흥행성 예측값
16	5	5.10	7682
17	8	8.16	5062
18	5	5.10	4765
5	9	9.18	3566
19	9	9.18	3043
10	7	7.14	2621
9	26	26.5	2232
14	8	8.16	2142
13	13	13.2	1781
15	8	8.16	718

표 11. 군집 1 영화 집단에 대한 이익도표

라. 군집 2 영화에 대한 의사결정나무분석

군집 2 집단에 대한 의사 결정 나무분석을 수행하기에 앞서 목표가 되는 종속변수와 분할기준으로 작용을 할 독립변수에 대한 기술 통계 값은 표 12 와 같다.

변수구분	변수명	평균	최대값	최솟값
종속변수	흥행도 (Nomal-Attendance)	4653	20155	503
독립변수	Happy	0.4704	0.74	0.11
	Surprise	0.2159	0.56	0.08
	Boring	0.0841	0.19	0.03
	Sad	0.1188	0.35	0.03
	Anger	0.0545	0.27	0.01
	Disgust	0.0370	0.30	0.00
	Fear	0.0182	0.22	0.00

표 12. 분석에 사용된 데이터의 특징(군집2)

분석에 사용된 군집1 집단에 대한 기술 통계량 값을 보면 흥행도는 종속 변수로써 평균값이 4653이며 7개의 대표 감정 어휘 값들이 독립변수로 사용되었고 Surprise, Happy의 감정어의 평균값이 다른 변수들에 비해 높은 것을 확인 할 수 있다.

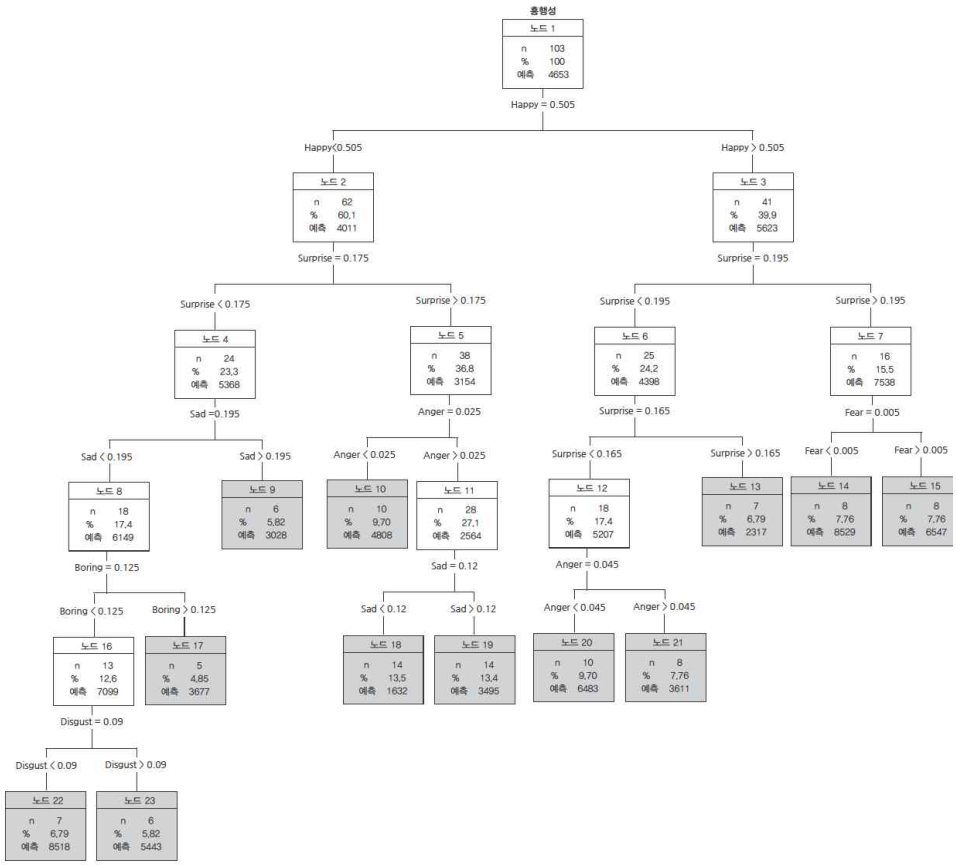

그림 10. 군집 2에 속한 영화에 대한 의사 결정 나무 분석

군집 2에 속한 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 영화 흥행성이 가장 높다고 예측된 집단에 대한 해석은 다음과 같다. 영화 흥행성이 8529가 되기 위해서는 Happy 0.505를 기준으로 최조 분리되어야 하며 Happy가 0.505이상 일 경우 Surprise 0.195를 기준으로 다시 분리된다. Surprise가 0.195이상 일 경우 마지막으로 Fear 0.005을 기준으로 분리되며 Fear가 0.005이하 일 때의 집단(N=8)에 대한 영화 흥행성의 예측 값은 8529로 높게 분류된다. 분석된 의사결정 나무분석 결과는 그림 10 과 같다.

의사결정 나무 분석의 결과는 이익도표를 통해 더 자세히 확인 할 수 있다. 군집 2에 속한 영화 집단에 대한 이익 도표 값은 표 13 과 같다.

노드번호	개수(N)	비율(%)	영화 흥행성 예측값
14	8	7.76	8529
22	7	6.79	8518
15	8	7.76	6547
20	10	9.70	6483
23	6	5.82	5443
10	10	9.70	4808
17	5	4.85	3677
21	8	7.76	3611
19	14	13.4	3495
9	6	5.82	3028
13	7	6.79	2317
18	14	13.5	1632

표 13. 군집 2 영화 집단에 대한 이익도표

마. 군집 3 영화에 대한 의사결정나무분석

군집 3 집단에 대한 의사 결정 나무분석을 수행하기에 앞서 목표가 되는 종속변수와 분할기준으로 작용을 할 독립변수에 대한 기술 통계 값은 표 14 와 같다.

변수구분	변수명	평균	최대값	최솟값
종속변수	흥행도 (Nomal-Attendance)	3970.1	15210.3	153.5
독립변수	Happy	0.3483	0.66	0.09
	Surprise	0.2595	0.65	0.08
	Boring	0.0956	0.25	0.02
	Sad	0.1409	0.48	0.04
	Anger	0.0669	0.20	0.01
	Disgust	0.0362	0.26	0.00
	Fear	0.0535	0.39	0.00

표 14. 분석에 사용된 데이터의 특징(군집3)

분석에 사용된 군집3 집단에 대한 기술 통계량 값을 보면 흥행도는 종속 변수로써 평균값이 3970.1이며 7개의 대표 감정 어휘 값들이 독립변수로 사용되었고 Surprise, Happy의 감정어의 평균값이 다른 변수들에 비해 높은 것을 확인할 수 있다.

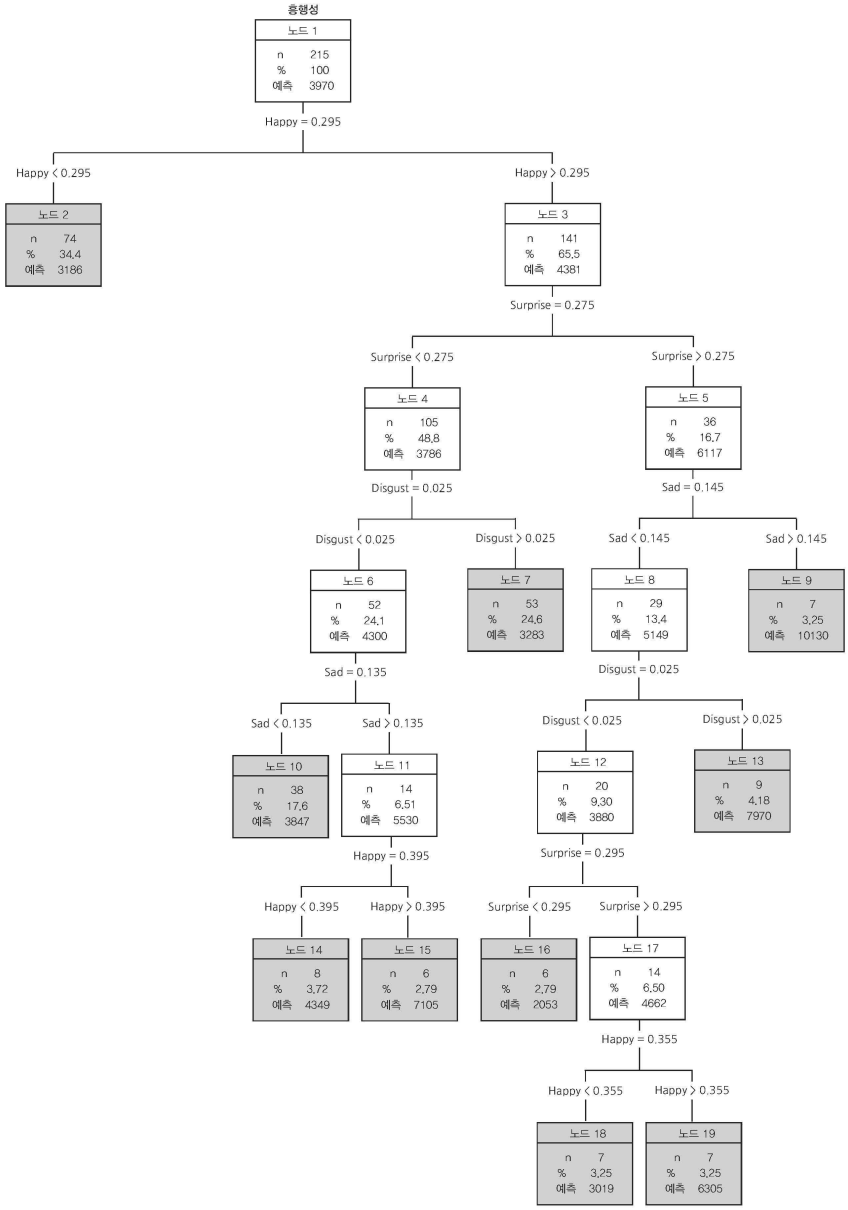

그림 11. 군집 3에 속한 영화에 대한 의사 결정 나무 분석

군집 3에 속한 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 영화 흥행성이 가장 높다고 예측된 집단에 대한 해석은 다음과 같다. 영화 흥행성이 10130이 되기 위해서는 Happy 0.295를 기준으로 최조 분리되어야 하며 Happy가 0.295이상 일 경우 Surprise 0.275를 기준으로 다시 분리된다. Surprise가 0.275이상 일 경우 마지막으로 Sad 0.145를 기준으로 분리되며 Sad가 0.145이하 일 때의 집단(N=7)에 대한 영화 흥행성의 예측값은 10130으로 높게 분류된다. 분석된 의사결정 나무분석 결과는 그림 11 과

같다.

의사결정 나무 분석의 결과는 이익도표를 통해 더 자세히 확인 할 수 있다. 군집 3 에 속한 영화 집단에 대한 이익 도표 값은 표 15 와 같다.

노드번호	개수(N)	비율(%)	영화 흥행성 예측값
9	7	3.25	10130
13	9	4.18	7970
15	6	2.79	7105
19	7	3.25	6305
14	8	3.72	4349
10	38	17.6	3847
7	53	24.6	3283
2	74	34.4	3186
18	7	3.25	3019
16	6	2.79	2053

표 15. 군집 3 영화 집단에 대한 이익도표

바. 군집 4 영화에 대한 의사결정나무분석

군집 4 집단에 대한 의사 결정 나무분석을 수행하기에 앞서 목표가 되는 종속변수와 분할기준으로 작용을 할 독립변수에 대한 기술 통계 값은 표 16 과 같다.

변수구분	변수명	평균	최대값	최솟값
종속변수	흥행도 (Nomal-Attendance)	4228	33590	35
독립변수	Happy	0.2584	0.67	0.05
	Surprise	0.3020	0.61	0.08
	Boring	0.0961	0.29	0.02
	Sad	0.1737	0.54	0.05
	Anger	0.0615	0.28	0.01
	Disgust	0.0419	0.30	0.00
	Fear	0.0659	0.61	0.00

표 16. 분석에 사용된 데이터의 특징(군집4)

분석에 사용된 군집4 집단에 대한 기술 통계량 값을 보면 흥행도는 종속 변수로써 평균값이 4228이며 대표 감정 어휘 값들이 독립변수로 사용되었고 Happy, Surprise의 평균값이 다른 변수들에 비해 높은 것을 확인 할 수 있다.

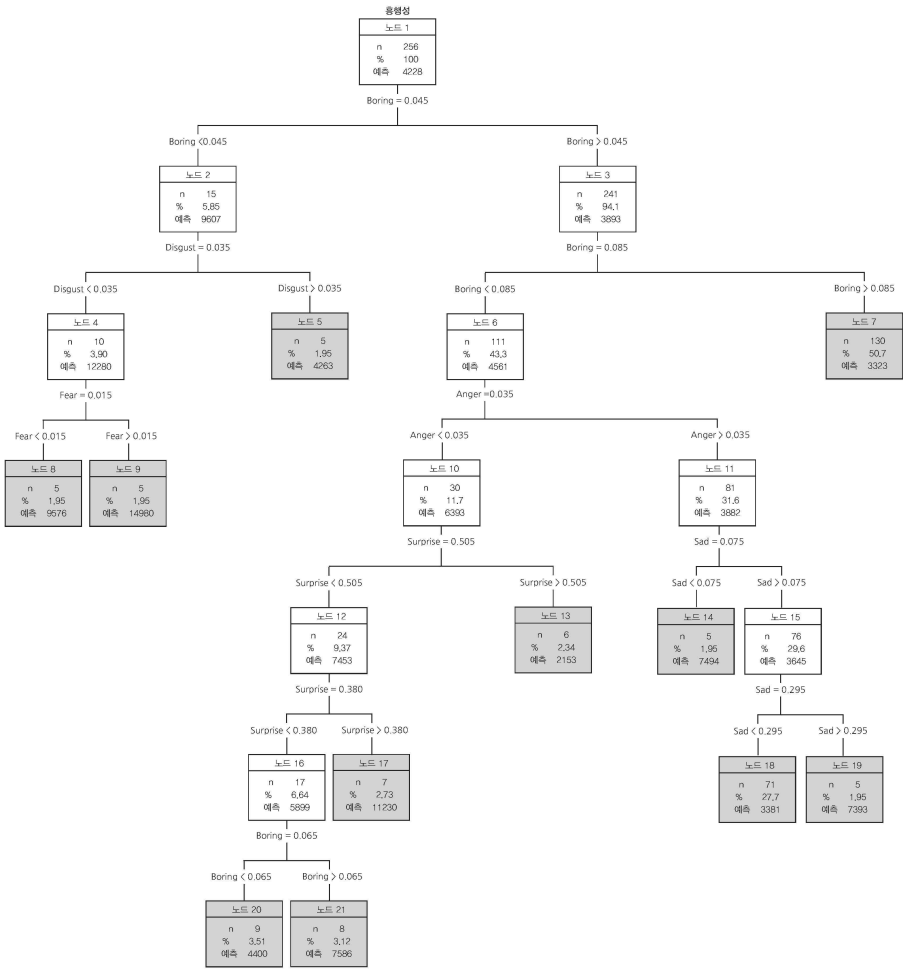

그림 12. 군집 4에 속한 영화에 대한 의사 결정 나무 분석

군집 4에 속한 영화 데이터 집단에 대한 최적분리는 Boring에 의해 최조 이지 분리 되었다. 영화 흥행성이 가장 높다고 예측된 집단에 대한 해석은 다음과 같다. 영화 흥행성이 14980이 되기 위해서는 Boring 0.045를 기준으로 최조 분리되어야 하며 Boring이 0.045이하 일 경우 Disgust 0.035를 기준으로 다시 분리된다. Disgust가 0.035이하 일 경우 마지막으로 Fear 0.015를 기준으로 분리되며 Fear가 0.015이상 일 때의 집단(N=5)에 대한 영화 흥행성의 예측 값은 14980으로 높게 분류된다. 분석된 의사결정 나무분석 결과는 그림 12 와 같다.

의사결정 나무 분석의 결과는 이익도표를 통해 더 자세히 확인 할 수 있다. 군집 4에 속한 영화 집단에 대한 이익 도표 값은 표 17 과 같다.

노드번호	개수(N)	비율(%)	영화 흥행성 예측값
9	5	1.95	14980
17	7	2.73	11230
8	5	1.95	9576
21	8	3.12	7586
14	5	1.95	7494
19	5	1.95	7393
20	9	3.51	4400
5	5	1.95	4263
18	71	27.7	3381
7	130	50.7	3323
13	6	2.34	2153

표 17. 군집 4 영화 집단에 대한 이익도표

사. 의사결정나무분석 결과에 대한 종합적인 해석

본 장에서는 전체집단과 4개의 군집화 된 집단에 대하여 영화의 흥행도가 높기 위해서는 어떠한 분할기준으로 집단이 분류되는지를 도출하고자 하였다. 각 집단 별로 의사결정나무분석을 수행하여 도출된 가장 높은 흥행도 예측 값을 보이는 최종마디에 대한 결과는 표 18 와 같다. 표 18에서 진한 색으로 색칠된 부분은 해당 값 이하인 분할기준이고 옅은 색으로 색칠된 부분은 해당 값 이상인 분할 기준이다.

군집	최종 노드	예측 값	개수	분할 1	분할 값	분할 2	분할 값	분할 3	분할 값	분할 4	분할 값
전체	14	17400	5	Happy	0.235	Anger	0.045	Sad	0.145	Boring	0.055
군집1	16	7680	5	Surprise	0.175	Boring	0.065	Happy	0.125	Fear	0.100
군집2	14	8529	8	Happy	0.505	Surprise	0.195	Fear	0.005		
군집3	9	10130	7	Happy	0.295	Surprise	0.275	Sad	0.145		
군집4	9	14980	5	Boring	0.045	Disgust	0.035	Fear	0.015		

표 18. 분석된 집단별 최종마디에 대한 결과

전체집단과 4개의 군집화 된 집단에 대하여 의사결정나무분석을 수행한 결과 가장 높은 영화 흥행도 예측 값이 도출된 집단은 전체영화에 대한 집단이었으며 흥행도 예측 값이 최대인 끝마디에 대한 분할기준은 Happy > 0.235 &

Anger < 0.045 & Sad > 0.145 & Boring < 0.055이었다. 이를 다시 말하면 전체 장르의 영화를 포함하는 전체 집단에서 흥행도의 예측 값이 최대가 되기 위해서는 행복한 감정, 슬픈 감정이 많이 느껴지고 화나는 감정과 지루한 감정이 낮게 느껴지는 영화가 높은 흥행도 값을 가질 수 있다고 해석 될 수 있다.
두 번째로 높은 영화 흥행도 예측 값이 도출된 집단은 군집4(Drama, SF, Crime, War)에 대한 집단이었으며 흥행도 예측 값이 최대인 끝마디에 대한 분할기준은 Boring < 0.045 & Disgust < 0.035 & Fear > 0.015이었다. 이를 다시 말하면 드라마, SF, 범죄, 전쟁 장르가 포함된 4번 군집에서 흥행도의 예측 값이 최대가 되기 위해서는 지루한 감정이 낮으며 역겨운 감정이 낮고 무서운 감정이 많이 느껴지는 영화가 높은 흥행도 값을 가질 수 있다고 해석 될 수 있다.
세 번째로 높은 영화 흥행도 예측 값이 도출된 집단은 군집3(Adventure, Meloromance, Action, Fantasy, Historicaldrama)에 대한 집단이었으며 흥행도 예측 값이 최대인 끝마디에 대한 분할기준은 Happy > 0.295 & Surprise > 0.275 & Sad > 0.145이었다. 이를 다시 말하면 모험, 멜로로멘스, 액션, 판타지, 사극 장르가 포함된 3번 군집에서 흥행도의 예측 값이 최대가 되기 위해서는 행복한 감정, 놀라운 감정, 슬픈 감정이 많이 느껴지는 영화가 높은 흥행도 값을 가질 수 있다고 해석 될 수 있다.
네 번째로 높은 영화 흥행도 예측 값이 도출된 집단은 군집2(Family, Comedy, Sports)에 대한 집단이었으며 흥행도 예측 값이 최대인 끝마디에 대한 분할기준은 Happy > 0.505 & Surprise > 0.195 & Fear < 0.005이었다. 이를 다시 말하면 가족, 코미디, 스포츠 장르가 포함된 2번 군집에서 흥행도의 예측 값이 최대가 되기 위해서는 행복한 감정, 놀라운 감정이 많이 느껴지고 무서운 감정이 낮게 느껴지는 영화가 높은 흥행도 값을 가질 수 있다고 해석 될 수 있다.
마지막으로 영화 흥행도 예측 값이 다른 집단 중에서 제일 낮은 집단은 군집1(Horror, Mystery, Thriller)이었으며 흥행도 예측 값이 최대인 끝마디에 대한 분할기준은 Surprise > 0.175 & Boring < 0.065 & Happy > 0.125 & Fear < 0.100이었다. 이를 다시 말하면 호러, 미스터리, 스릴러 장르가 포함된 1번 군집에서 흥행도의 예측 값이 최대가 되기 위해서는 행복한 감정, 놀라운 감정이 많이 느껴지고 무서운 감정과 지루한 감정이 낮게 느껴지는 영화가 높은 흥행도 값을 가질 수 있다고 해석 될 수 있다.

Ⅴ. 시각화 분석 및 검증

1. Parallel coordinates의 개념

본 연구의 데이터처럼 다 변량으로 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[33]. 이 방법은 1985년 Inselberg. A. 가 구체적으로 제안하였고 최근까지 다양한 학문 영역에서 사용되고 있다. Inselberg. A.의 연구에 따르면 Parallel coordinate는 각 변수가 대부분 라인이 평행일 때 두 차원 사이에 유사한 관계가 형성된다고 해석할 수 있으며, 대부분의 라인이 교차할 때는 상이한 관계가 형성된다고 해석한다[34].

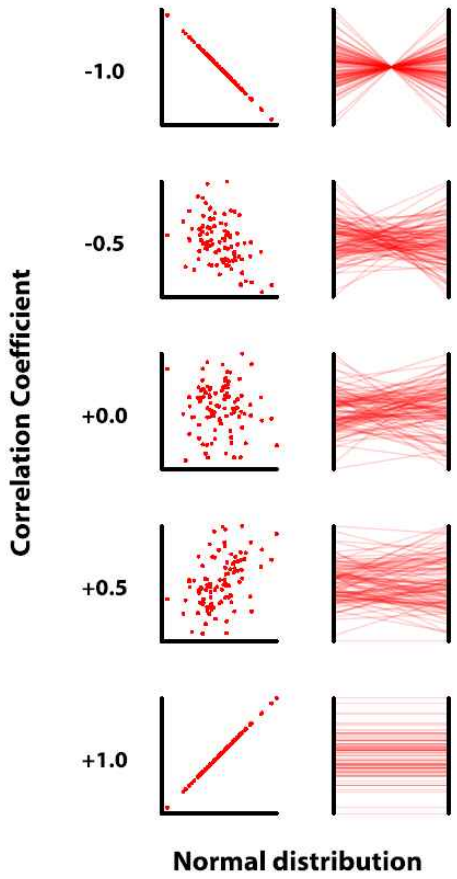

그림 13. 분포에 따른 Parallel coordinates

[33] Rick Walker, Philip A. Legg, Serban Pop, Zhao Geng, Robert S. Laramee,

또한 본 연구에서는 다 변량 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 인터렉션 기능들이 있다.

2. Parallel coordinates의 기능

본 연구에서는 다 변량으로 이루어진 데이터를 분석하고 통계분석 결과를 검증하기 위해 기존의 Parallel coordinates 시각화 방법에 다양한 기능을 추가하였다. 해당 시각화는 http://stat34.github.io/parallel_coordinate_final/ 에서 사용할 수 있으며, 연구에 사용된 Parallel coordinates 시각화의 기능은 다음과 같다.

가. 번들링(Bundling)

Parallel coordinates 시각화는 일반적으로 축이 변경 될 때 데이터의 연결 표현을 직선으로 표현하고 있다. 하지만 직선으로 데이터 사이의 연결을 표현할 경우 데이터의 양이 많으면 축이 전혀 보이지 않고 데이터 사이의 패턴 또한 발견하기 어렵다. 본 연구에서는 그림 14 의 오른쪽과 같이 데이터 사이의 연결 표현에 번들링 기능을 추가 하였으며 번들링 기능을 통해 데이터 사이의 연결을 표현하면 해당 축의 데이터들을 일정 수준으로 묶어서 표현하기 때문에 데이터들이 군집화 되는 경향을 쉽게 확인 할 수 있다.

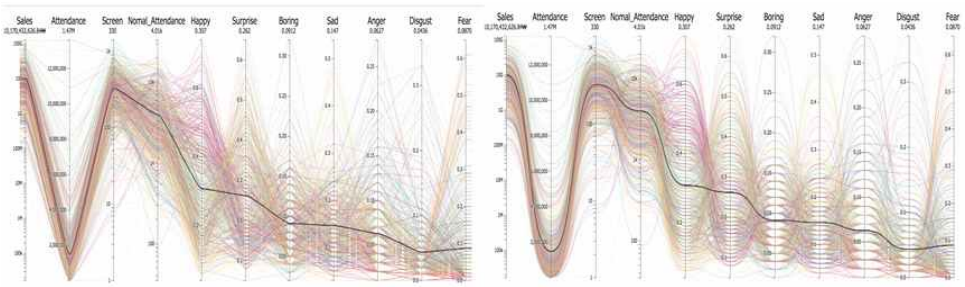

그림 14. (왼쪽) 일반적인 Parallel coordinates (오른쪽) 번들링 기능을 추가한 Parallel coordinates

Jonathan C. Roberts, "Force-Directed Parallel Coordinates", 17th International Conference on Information Visualisation, p.39, 2013.

34 Inselberg, A, The plane with Parallel coordinates, The Visual Computer, p. 79, 1985.

나. 축(Axes)

일반적인 Parallel coordinates 시각화는 데이터 변수를 축으로 설정하고 데이터 변수 사이에서 변화하는 데이터들의 패턴을 확인하는데 매우 용이하다. 본 연구에서는 Parallel coordinates에 그림 15 와 같이 데이터 변수 축의 삭제 기능과 축의 순서를 이동시키는 기능을 추가하여 분석을 용이하게 하였다.

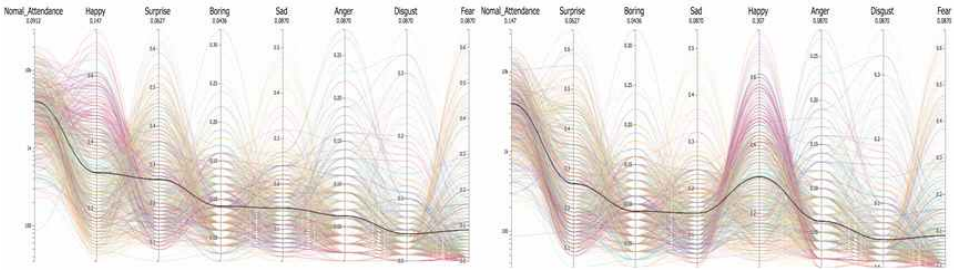

그림 15. (왼쪽) 기본적인 데이터 축의 순서 (오른쪽) Happy의 데이터 변수 축 순서 변경

다. 색상(Colour)

장르별로 비교 분석을 하기 위해서는 영화의 장르별로 구분 할 수 있는 기능이 필요한데, 본 연구에서는 그림 16 과 같이 영화가 속하는 주요 장르에 따라 line 그래프의 색상을 다르게 지정하여 사용자가 영화의 장르별로 데이터를 구분 할 수 있게 하였다.

그림 16. 영화의 장르별로 지정된 색상

라. 기술 통계(Descriptive statistic)

일반적으로 시각화 분석에서 사용되는 Parallel coordinates는 데이터 패턴, 데이터 변수 축에서 발생하는 군집화, 변수 축 사이의 직선 기울기 등 시각적으로 확인이 가능한 부분만으로 해석을 해야 한다. 본 연구에서는 다양한 관점으로 분석을 시행하기 위해 그림 17 과 그림 18 과 같이 선택된 데이터의 평

균을 나타내는 평균선, 선택된 각각의 데이터 변수 축의 평균 값, 선택된 영화 수의 합계를 내타내는 기능을 추가하였다.

그림 17. 선택된 데이터 변수들의 평균값과 영화 수의 합계

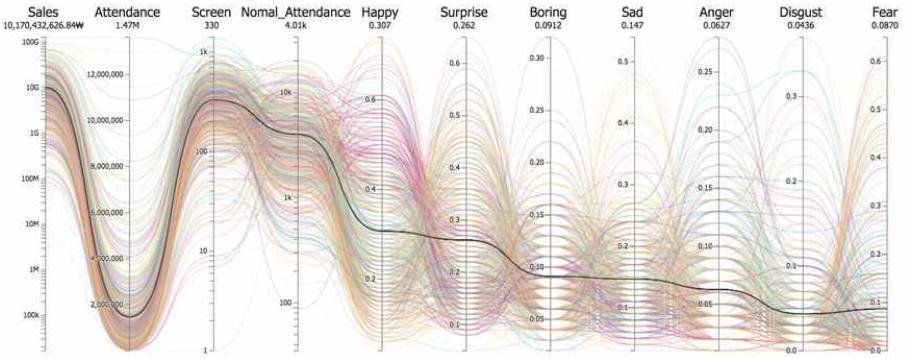

그림 18. 선택된 데이터 변수들의 평균값과 평균 선 (굵은 line 그래프)

마. 데이터 선택(Data Selection)

영화의 장르 별로 대표 감정 어휘의 분포를 탐색하고 영화 흥행과 감정 어휘 사이의 관계를 파악하기 위해서는 장르 별 데이터의 패턴 비교, 분포 확인, 조건에 따른 패턴 변화 등을 확인하여야 한다. 본 연구에서는 분석을 용이하기 위해 그림 19 와 그림 20과 같이 장르 선택 기능, 영화 검색 기능, 조건에 따른 데이터 필터링 기능, 하이라이트 기능 등을 추가하여 분석을 용이하게 하였다.

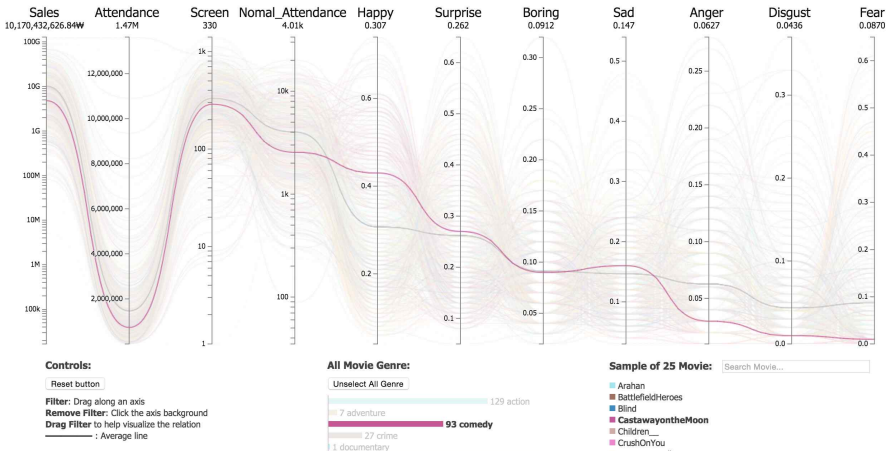

그림 19. 김씨 표류기(Castaway on the Moon)에 대한 하이라이트

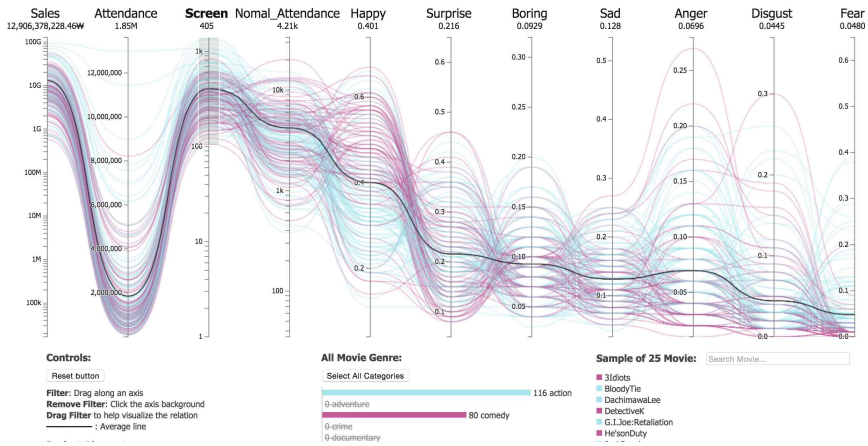

그림 20. 장르가 액션 & 코미디이고 상영 스크린 수가 100개 이상

바. 제거된 데이터 표현

선행된 통계적인 분석 방법 중 의사결정나무분석 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 21, 그림 22 와 같다.

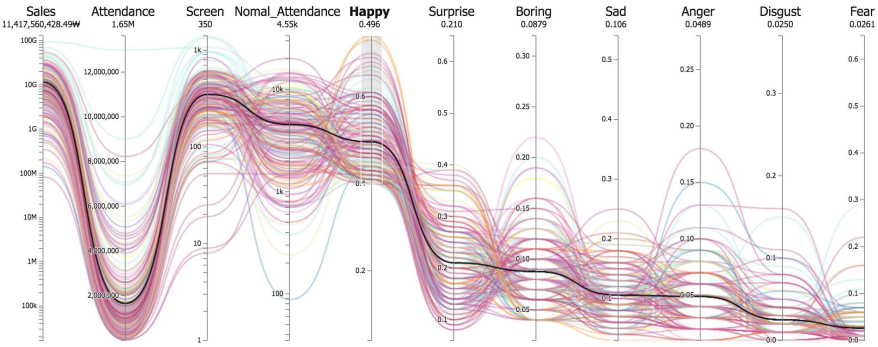

그림 21. 선택되지 않은 데이터가 제거되는 화면

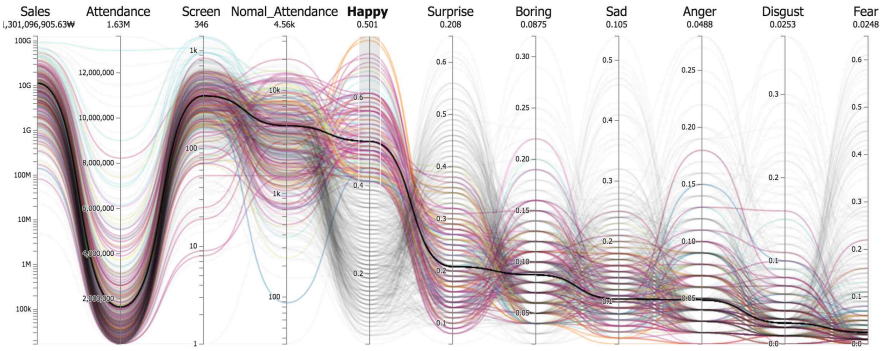

그림 22. 선택되지 않은 데이터를 표현하는 화면

3. Parallel coordinates를 활용한 분석

본 장에서는 7개의 대표 감정 어휘, 영화 티켓 판매액, 영화 관람 관객 수, 상영 스크린 수, 한 스크린 당 영화 관람 관객 수, 영화의 장르, 영화의 영문 이름 등으로 이루어진 최종 데이터를 활용하여 시각화 분석을 수행하고 앞서 선행되었던 통계적적인 분석 결과를 검증하고자 한다.

가. 영화 장르 별 흥행도의 분포와 대표 감정 어휘의 분포

본 연구에 사용된 672개의 영화의 대표 장르는 드라마(25.9%), 액션(19.2%), 코미디(13.8%), 멜로 로맨스(7.6%), SF(7.3%)등의 순서이며 데이터의 개수가 15개 이상인 영화의 장르 별 흥행도와 감정 어휘 분포는 표 19 와 같다.

대표 장르	흥행도	Happy	Surprise	Boring	Sad	Anger	Disgust	Fear
드라마(174)	4,270	0.249	0.333	0.0935	0.186	0.054	0.035	0.0476
액션(129)	4,060	0.351	0.233	0.0926	0.13	0.0774	0.0455	0.0715
코미디(92)	4,860	0.465	0.213	0.0849	0.122	0.0561	0.0397	0.0177
멜로로맨스(52)	3,590	0.325	0.330	0.102	0.161	0.0438	0.021	0.019
SF(49)	4,130	0.274	0.227	0.102	0.163	0.0833	0.0504	0.102
호러(41)	2,480	0.120	0.176	0.0573	0.113	0.0495	0.080	0.404
스릴러(40)	3,520	0.207	0.253	0.0898	0.118	0.0828	0.0688	0.182
범죄(27)	3,680	0.283	0.250	0.101	0.121	0.0715	0.0722	0.103
판타지(18)	3,310	0.368	0.240	0.104	0.159	0.0728	0.0239	0.0339
미스터리(17)	2,480	0.225	0.217	0.0912	0.123	0.0506	0.0347	0.259

표 19. 데이터의 개수가 15개 이상인 장르 별 감정 어휘 분포(평균 값)

영화의 흥행도가 제일 높은 장르인 코미디의 경우, 한 스크린 당 평균 누적 관객수는 4,860명 이었고 감정어의 분포는 그림 23 과 같이 'Happy'(0.465), 'Surprise'(0.213), 'Sad'(0.122), 'Boring'(0.0849), 'Anger'(0.0561), 'Disgust'(0.0397), 'Fear'(0.0177)순이었으며 평균적으로 행복한 감정이 다른 감정에 비해 많이 느껴진다는 결과를 Parallel coordinates를 통해 확인 할 수 있다.

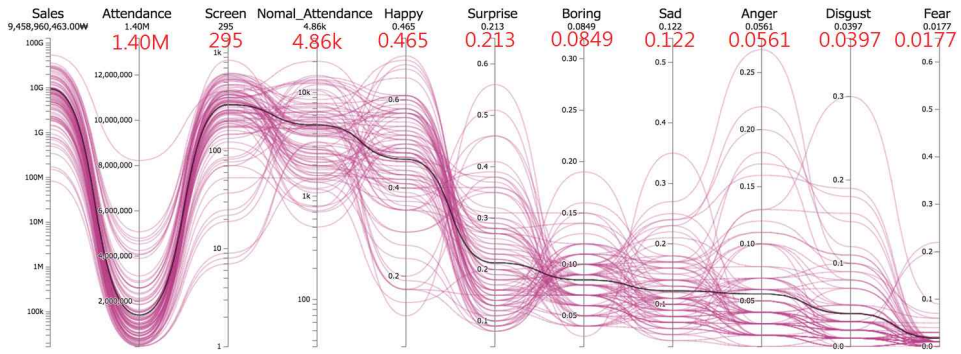

그림 23. 영화의 대표 장르가 코미디인 영화에 대한 감정 어휘 분포

다음으로 영화의 흥행도가 제일 낮은 장르인 호러의 경우, 한 스크린 당 평균 누적 관객 수는 2,480명 이었고 감정어의 분포는 그림 24 와 같이 'Fear'(0.404), 'Surprise'(0.176), 'Happy'(0.120), 'Sad'(0.113), 'Disgust'(0.0800), 'Boring'(0.0573), 'Anger'(0.0495)순이었으며 다른 감정에 비해 무서운 감정이 많이 느껴진다는 결과를 Parallel coordinates를 통해 확인 할 수 있다.

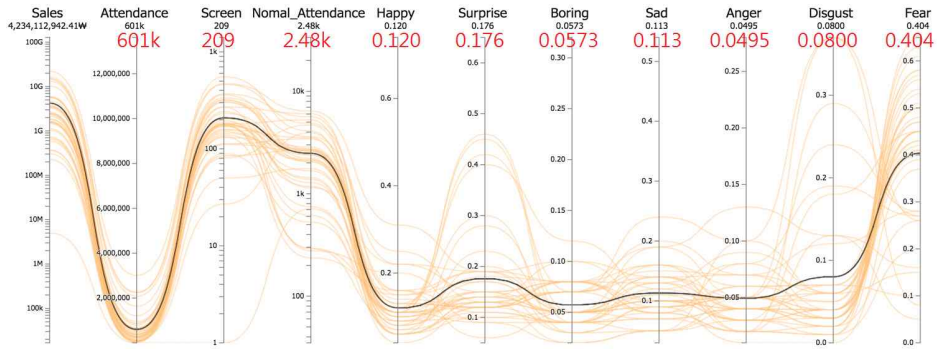

그림 24. 영화의 대표 장르가 호러인 영화에 대한 감정 어휘 분포

나. 영화의 대표 감정 어휘 사이의 상관관계

본 연구에서는 감정 어휘를 36개로 세분화 하였으며 이를 다시 'Happy', 'Surprise', 'Boring', 'Sad', 'Anger', 'Disgust', 'Fear'로 군집화 하였다. 장르 구분 없이 전체 영화를 대상으로 상관계수(Correlation) 공식을 활용해 대표 감정 어휘 별 상관관계를 분석한 결과는 표 20 과 같으며 모든 변수들은 $P < 0.05$로 선형 연관성이 있었다.

	Happy	Surprise	Boring	Sad	Anger	Disgust	Fear
Happy	1.00	-0.286	0.016	-0.444	-0.175	-0.300	-0.466
Surprise	-0.286	1.00	-0.136	0.042	-0.407	-0.338	-0.300
Boring	0.016	-0.136	1.00	0.052	0.068	-0.173	-0.205
Sad	-0.444	0.042	0.052	1.00	0.202	0.045	-0.215
Anger	-0.175	-0.407	0.068	0.202	1.00	0.514	-0.097
Disgust	-0.300	-0.338	-0.173	0.045	0.514	1.00	0.122
Fear	-0.466	-0.300	-0.205	-0.215	-0.097	0.122	1.00

표 20. 전체 영화의 대표 감정 어휘 별 상관관계 표

전체 영화를 대상으로 대표 감정 어휘 별 상관관계를 분석한 결과 상관관계가 정(正)의 방향으로 가장 높은 감정 어휘는 'Disgust'와 'Anger'로 상관관계는 0.514이다. 이를 Parallel coordinates를 통해 확인 한 결과는 다음과 같다.

‘Disgust’에 대한 구간 값이 0.1이하인 경우 ‘Disgust’와 ‘Anger’의 감정 어휘 값은 0.0295와 0.0560, ‘Disgust’에 대한 구간 값이 0.1부터 0.3인 경우 ‘Disgust’와 ‘Anger’의 감정 어휘 값은 0.160와 0.123, ‘Disgust’에 대한 구간 값이 0.3이상인 경우 ‘Disgust’와 ‘Anger’의 감정 어휘 값은 0.350와 0.143이었다. 이는 ‘Disgust’의 값이 높아질 때 ‘Anger’의 값도 같이 높아진다고 볼 수 있으며 분석 결과는 그림 25 와 같다.

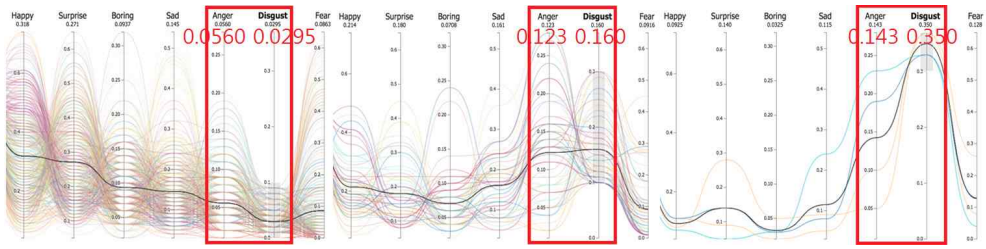

그림 25. Disgust값이 0.1이하(왼쪽), 0.1이상 0.3이하(가운데), 0.3이상(오른쪽)일 때의 분석 결과

다음으로 상관관계가 부(不)의 방향으로 가장 높은 감정 어휘는 ‘Happy’와 ‘Sad’로 상관관계는 -0.444이다. 이를 Parallel coordinates를 통해 확인 한 결과는 다음과 같다. ‘Sad’에 대한 구간 값이 0.175이하인 경우 ‘Happy’와 ‘Sad’의 감정 어휘 값은 0.341와 0.110, ‘Sad’에 대한 구간 값이 0.175부터 0.35인 경우 ‘Happy’와 ‘Sad’의 감정 어휘 값은 0.226와 0.224, ‘Sad’에 대한 구간 값이 0.35이상인 경우 ‘Happy’와 ‘Sad’의 감정 어휘 값은 0.110와 0.433이었다. 이는 ‘Sad’의 값이 높아질 때 ‘Happy’의 값은 낮아진다고 볼 수 있으며 분석 결과는 그림 26 과 같다.

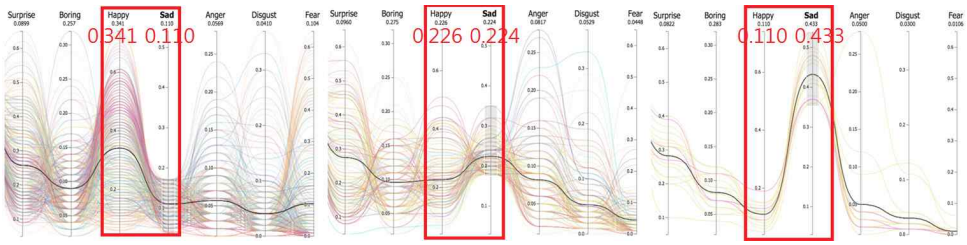

그림 26. Sad값이 0.175이하(왼쪽), 0.175이상 0.35이하(가운데), 0.35이상(오른쪽)일 때의 분석 결과

마지막으로 모든 영화 장르에서 값이 높게 나왔던 ‘Happy’와 ‘Surprise’의 상관관계를 분석한 결과 부(不)의 방향으로 상관관계(-0.286)가 있는 것을 확인하였다. 이를 Parallel coordinates를 통해 확인 한 결과는 다음과 같다. ‘Surprise’에 대한 구간 값이 0.2이하인 경우 ‘Happy’와 ‘Surprise’의 감정 어휘 값은 0.330과 0.144, ‘Surprise’에 대한 구간 값이 0.2부터 0.4인 경우

‘Happy’와 ‘Surprise’의 감정 어휘 값은 0.324와 0.282, ‘Surprise’에 대한 구간 값이 0.4이상인 경우 ‘Happy’와 ‘Surprise’의 감정 어휘 값은 0.203와 0.468이었다. 이는 ‘Surprise’의 값이 높아질 때 ‘Happy’의 값은 낮아진다고 볼 수 있으며 분석 결과는 그림 27 과 같다.

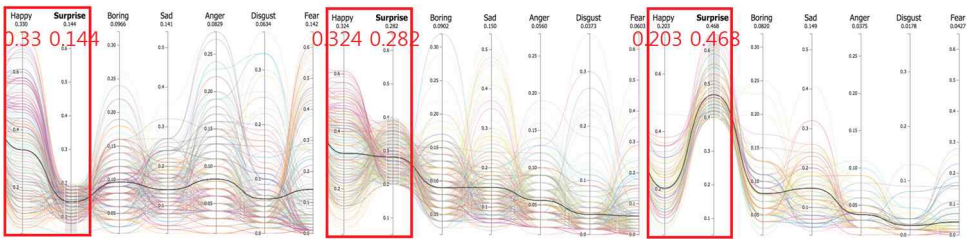

그림 27. Surprise값이 0.2이하(왼쪽), 0.2이상 0.4이하(가운데), 0.4이상(오른쪽)일 때의 분석 결과

4. 통계분석 결과에 대한 시각화 검증

본 장에서는 선행된 의사결정나무분석 예측모형에 대해 시각화 분석 방법을 활용하여 검증하고자 한다. 의사결정나무분석을 활용하여 도출된 예측모형은 비슷한 감정을 느끼는 장르별로 군집 화된 집단과 전체 영화집단에 대하여 어떠한 감정이 느껴질 경우 흥행도의 예측 값이 최고가 될 수 있는지를 제안하는데 매우 유용하게 활용될 수 있다. 하지만 의사결정나무분석의 경우 패턴인식(Pattern recognition) 혹은 기계학습(Macine learning)으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 분석된 결과 외에는 일반적인 사용자가 유동적으로 분석 과정을 볼 수 없다는 단점이 있다[35]. 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 제안하고자 한다.

가. 전체 영화에 대한 의사결정나무분석 및 시각화 검증

전체 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 의사결정나무 분석을 통해 영화 흥행성의 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 흥행도 값이 예측된 노드 14에 대한 데이터 분할 과정을 Parallel coordinates시각화 분석 방법을 통해 검증하고자 한다. 노드 14에 대해서 분할기준이 적용되지 않은 전체 값은 그림 28 과 같다.

35 Soon Tee Teoh, KwanLiu Ma. p. 667, 2003.

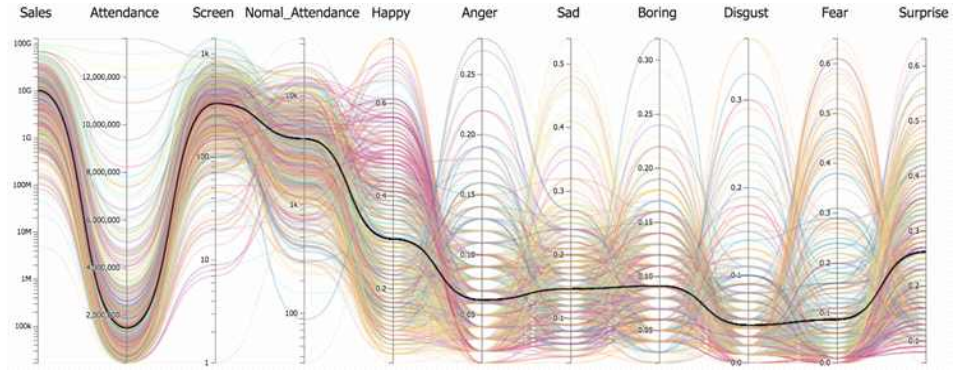

그림 28. 분할기준이 적용되기 전의 Parallel coordinates시각화

다음으로 Happy의 값이 0.235이상 일 때와 아닐 때로 최초 분할되었으며 Happy의 값이 0.235이상인 노드에 대한 시각화 결과는 그림 29 와 같다.

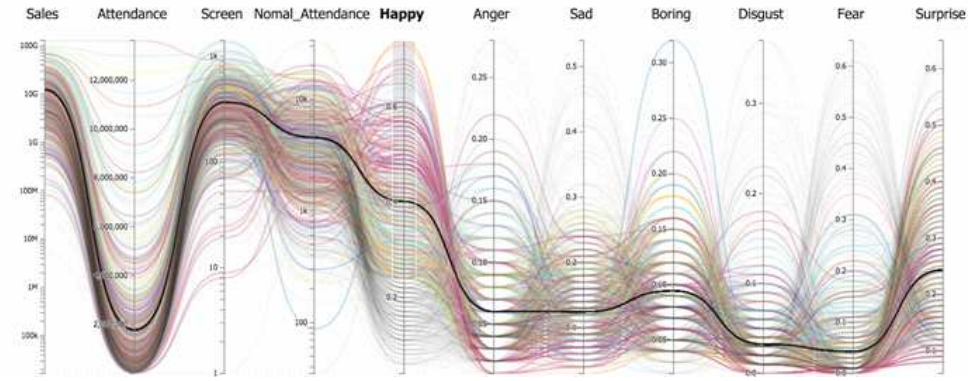

그림 29. Happy > 0.235이 적용된 시각화 결과

Happy를 기준으로 0.235이상인 데이터를 선택하였을 때 Happy의 값이 대체적으로 낮았던 드라마 장르의 영화와 호러장르의 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 30 과 같다.

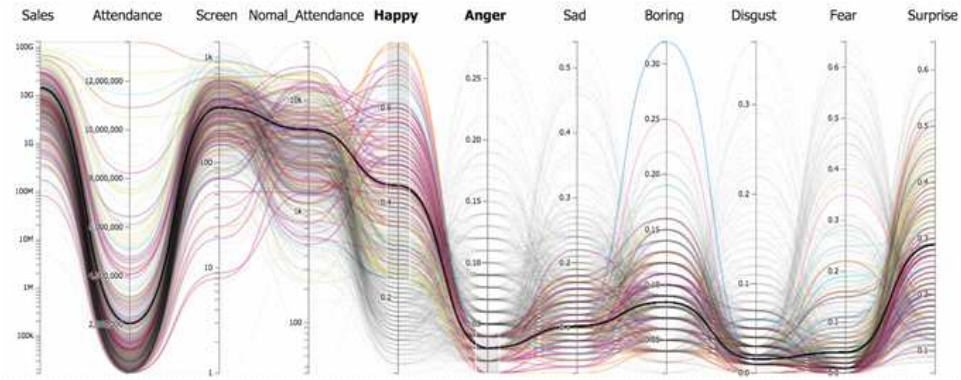

그림 30. Happy > 0.235 & Anger < 0.045이 적용된 시각화 결과

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 31 과 같다.

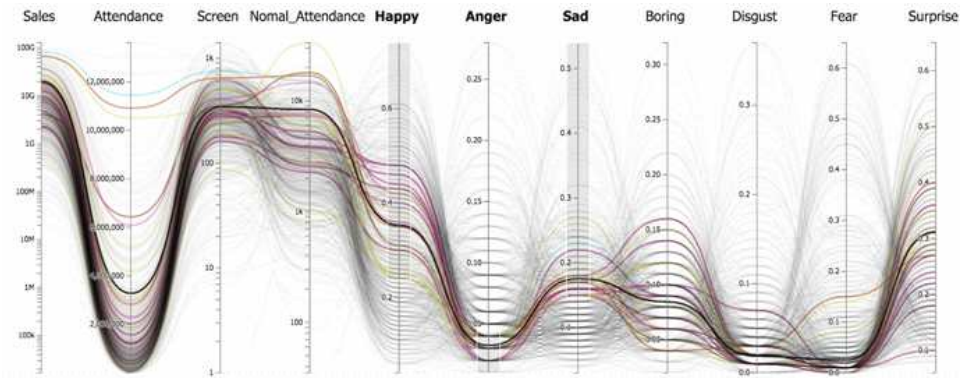

그림 31. Happy > 0.235 & Anger < 0.045 & Sad > 0.145이 적용된 시각화 결과

Happy는 0.235이상, Anger는 0.045이하, Sad는 0.145이상인 데이터를 선택하였을 때 코미디와 드라마장르가 각각 8편과 10편으로 선택된 23편의 영화중 가장 많은 비율을 차지하였다. 마지막으로 적용된 분할 기준은 Boring이며 Boring값이 0.055이하 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각화 결과는 그림 32, 그림 33 과 같다.

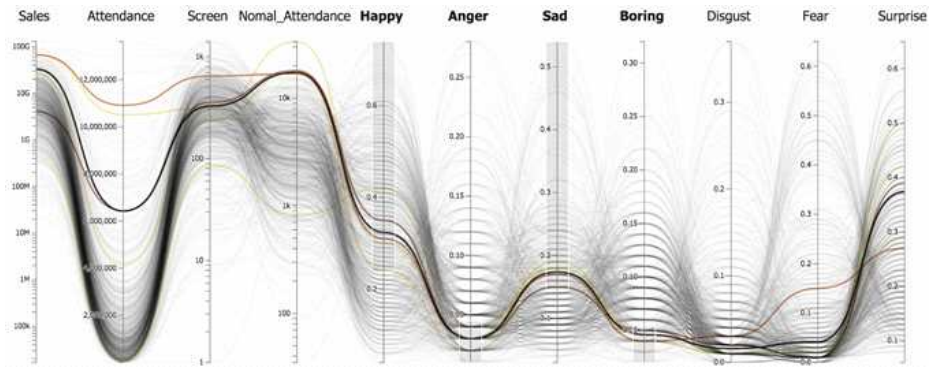

그림 32. 노드 14 에 대한 최종 Parallel coordinates

그림 33. 노드 14 에 최종 포함된 영화 정보

의사결정나무 분석을 통해 영화 흥행성의 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. 분할 과정에서 최종 까지 선택된 라인의 색상을 진하게 하여 표현한 전체 영화에 대한 시각화 모습은 그림 34 와 같다.

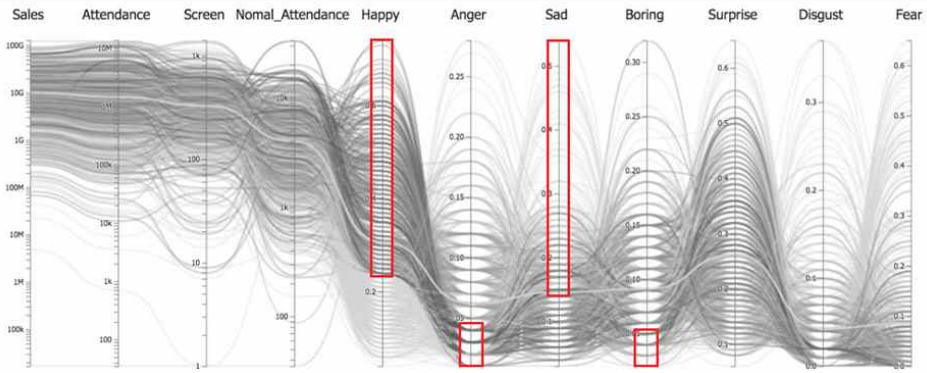

그림 34. 전체 영화에 대한 최종 Parallel coordinates

나. 군집 1 영화에 대한 의사결정나무분석 및 시각화 검증

군집1(Horror, Mystery, Thriller) 에 속한 영화 데이터 집단에 대한 최적분리는 Surprise에 의해 최조 이지 분리 되었다. 의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 16에 해당하는 집단의 분할 규칙은 Surprise > 0.175 & Boring < 0.065 & Happy > 0.125 & Fear < 0.100의 순서로 4번 분할 된 것을 확인 할 수 있다. 이번 장에서는 군집1 에서 최대의 흥행도 값이 예측된 노드 16에 대한 데이터 분할 과정을 Parallel coordinates시각화 분석 방법을 통해 검증하고자 한다. 노드 16에 대해서 분할기준이 적용되지 않은 전체 값은 그림 35 와 같다.

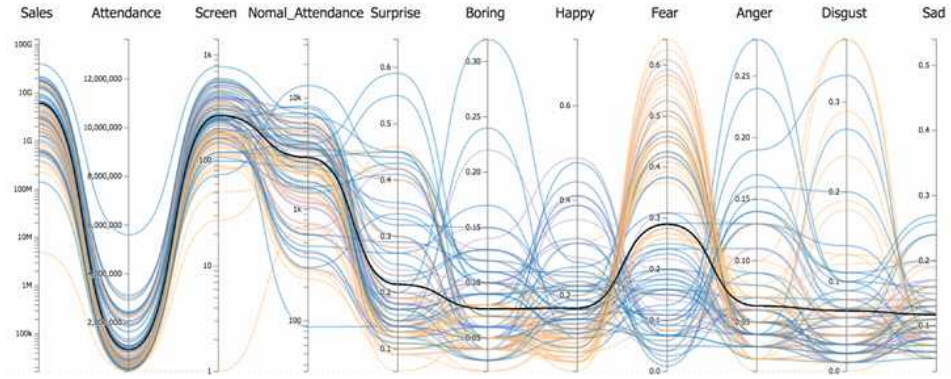

그림 35. 분할기준이 적용되기 전의 Parallel coordinates시각화

다음으로 Surprise의 값이 0.175이상 일 때와 아닐 때로 최초 분할되었으며 Surprise의 값이 0.175이상인 노드에 대한 시각화 결과는 그림 36 과 같다.

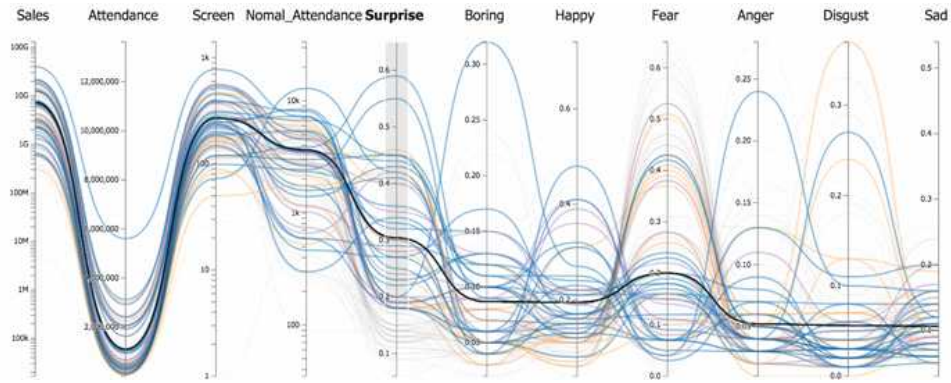

그림 36. Surprise > 0.175이 적용된 시각화 결과

Surprise를 기준으로 0.175이상인 데이터를 선택하였을 때 Surprise의 값이 낮았던 호러 장르의 영화가 41에서 13으로 대폭 감소하였다. 두 번째로 적용된 분할 기준은 Boring이며 Boring값이 0.065이하 일 때와 아닐 때로 분류된 시각화 결과는 그림 37 과 같다.

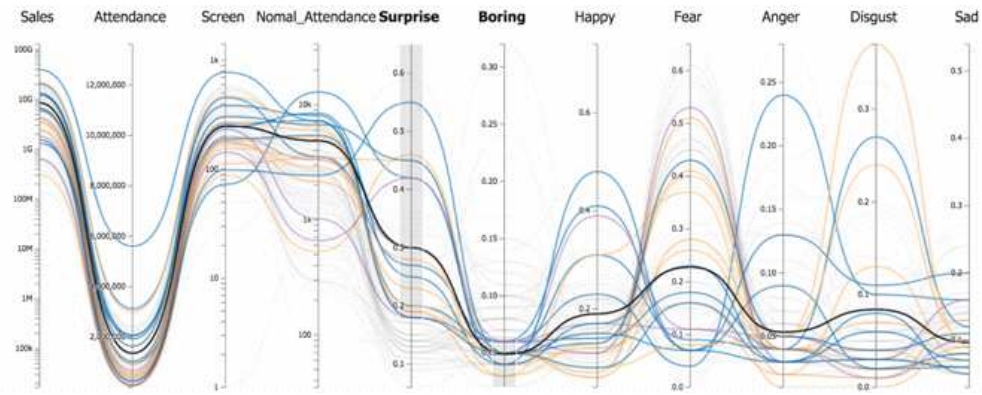

그림 37. Surprise > 0.175 & Boring < 0.065이 적용된 시각화 결과

Surprise는 0.175이상이고 Boring는 0.065이하인 데이터를 선택하였을 때 스릴러, 미스터리 장르의 영화가 대폭 감소하였다. 스릴러의 경우 24에서 9로, 미스터리의 경우 10에서 3으로 감소한 것을 확인하였다. 세 번째로 적용된 분할 기준은 Happy이며 Happy값이 0.125이상 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각화 결과는 그림 38 과 같다.

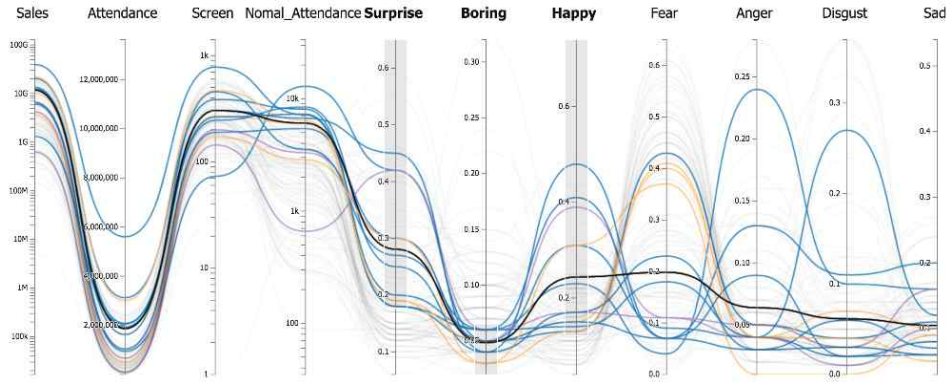

그림 38. Surprise > 0.175 & Boring < 0.065 & Happy > 0.125이 적용된 시각화 결과

Surprise는 0.175이상, Boring은 0.065이하, Happy는 0.0.125이상인 데이터를 선택하였을 때 호러와 스릴러 장르가 각각 5편과 8편으로 선택된 18편의 영화중 가장 많은 비율을 차지하였다. 마지막으로 적용된 분할 기준은 Fear이며 Fear값이 0.100이하 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각화 결과는 그림 39, 그림 40 과 같다.

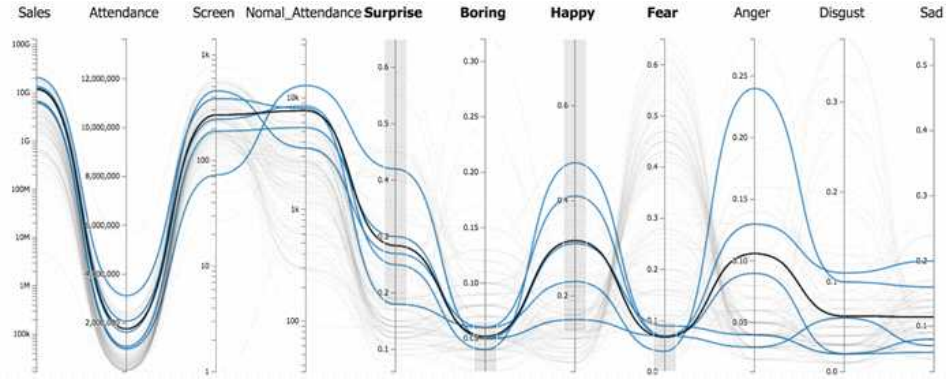

그림 39. 노드 16 에 대한 최종 Parallel coordinates

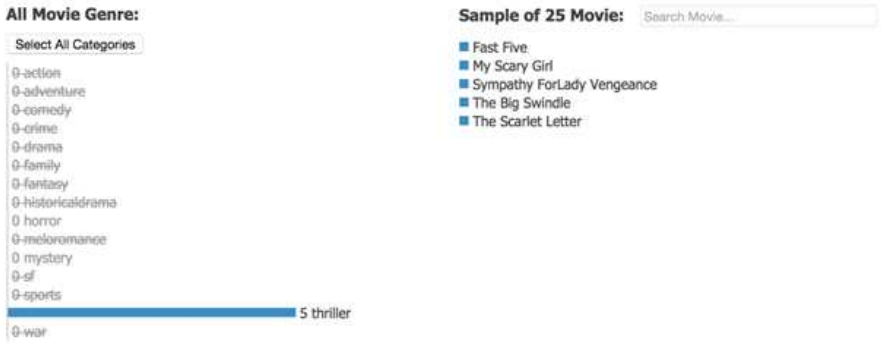

그림 40. 노드 16 에 최종 포함된 영화 정보

의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 16에 대해서 Parallel coordinates시각화 분석 방법을 사용하여 분석 한 결과, 노드 16 집단에 포함된 영화는 Fast Five, My Scary Girl, Sympathy ForLady Vengeance, The Big Swindle, The Scarlet Letter이었으며 영화의 장르는 모두 Thriller 인 것을 확인 할 수 있었다. 분할 과정에서 최종 까지 선택된 라인의 색상을 진하게 하여 표현한 전체 영화에 대한 시각화 모습은 그림 41 과 같다.

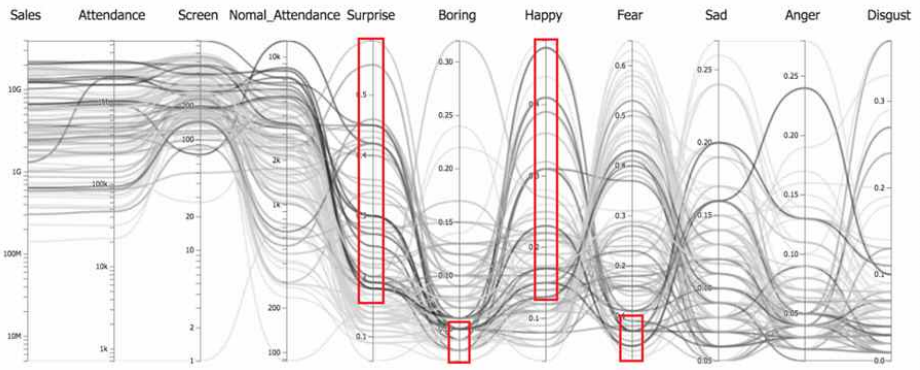

그림 41. 군집1에 대한 최종 Parallel coordinates

다. 군집 2 영화에 대한 의사결정나무분석 및 시각화 검증

군집2(Family, Comedy, Sports)에 속한 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 14에 해당하는 집단의 분할 규칙은 Happy > 0.505 & Surprise > 0.195 & Fear < 0.005의 순서로 3번 분할 된 것을 확인 할 수 있다. 이번 장에서는 군집2 에서 최대의 흥행도 값이 예측된 노드 14에 대한 데이터 분할 과정을 Parallel coordinates시각화 분석 방법을 통해 검증하고자 한다. 노드 14에 대해서 분할기준이 적용되지 않은 전체 값은 그림 42 와 같다.

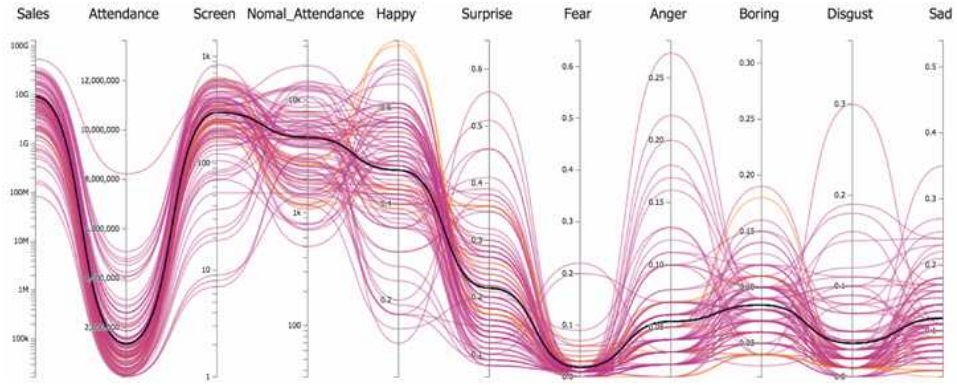

그림 42. 분할기준이 적용되기 전의 Parallel coordinates시각화

다음으로 Happy의 값이 0.505이상 일 때와 아닐 때로 최초 분할되었으며 Happy의 값이 0.505이상인 노드에 대한 시각화 결과는 그림 43 과 같다.

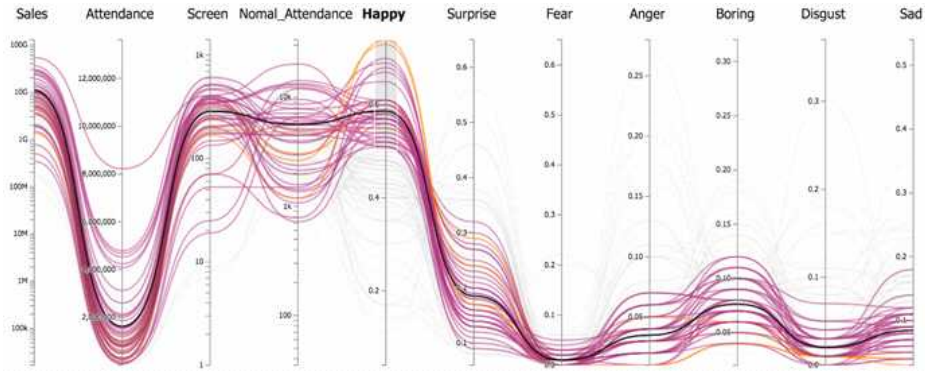

그림 43. Happy > 0.505 이 적용된 시각화 결과

Happy를 기준으로 0.505이상인 데이터를 선택하였을 때 Happy의 값이 낮았던 코미디 장르의 영화가 93에서 37로 대폭 감소하였으며 스포츠 장르의 영화는 모두 제거되었다. 두 번째로 적용된 분할 기준은 Surprise이며 Surprise값이 0.195이상 일 때와 아닐 때로 분류된 시각화 결과는 그림 44 와 같다.

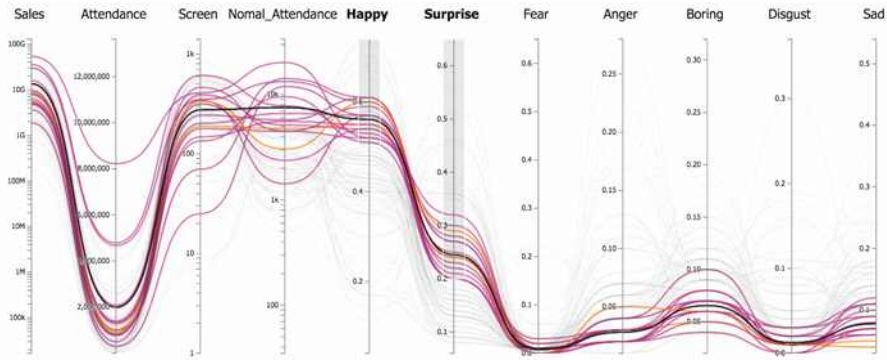

그림 44. Happy > 0.505 & Surprise > 0.195 이 적용된 시각화 결과

Happy는 0.505이상이고 Surprise는 0.195이상인 데이터를 선택하였을 때 코미디 장르가 14편으로 선택된 16편의 영화중 가장 높은 비율을 차지하였다. 마지막으로 적용된 분할 기준은 Fear이며 Fear값이 0.005이하 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각화 결과는 그림 45, 그림 46 과 같다.

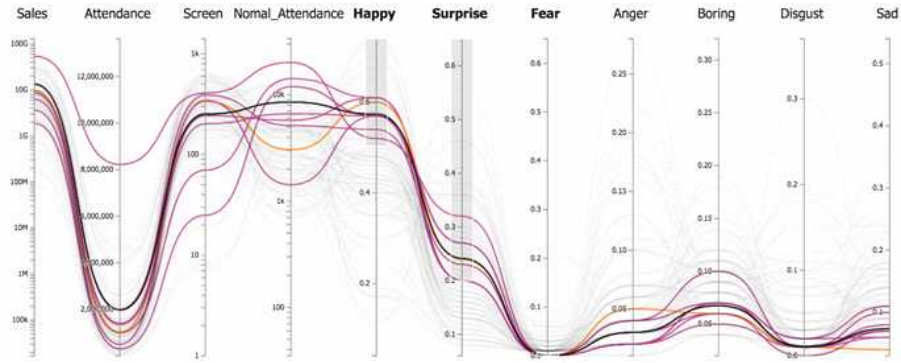

그림 45. 노드 14 에 대한 최종 Parallel coordinates

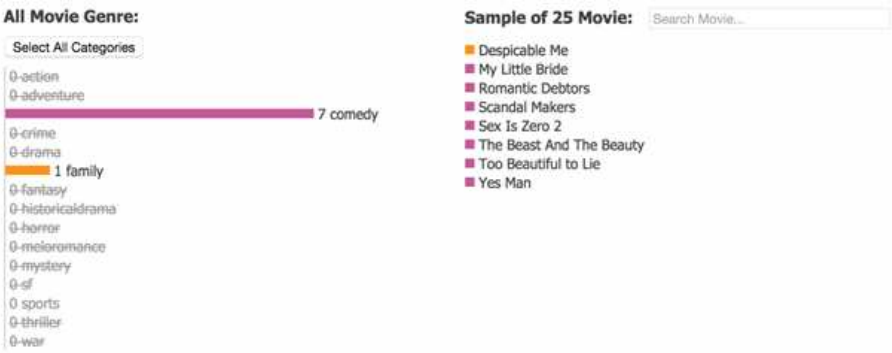

그림 46. 노드 14 에 최종 포함된 영화 정보

의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 14에 대해서 Parallel coordinates시각화 분석 방법을 사용하여 분석 한 결과, 노드 14 집단에 포함된 영화는 Despicable Me, My Little Bride, Romantic Debtors, Scandal Makers, Sex Is Zero 2, The Beast And The Beauty, Too Beautiful to Lie, Yes Man이었으며 영화의 장르는 Comedy가 7, Family가 1 포함된 것을 확인 할 수 있었다. 분할 과정에서 최종 까지 선택된 라인의 색상을 진하게 하여 표현한 전체 영화에 대한 시각화 모습은 그림 47 와 같다.

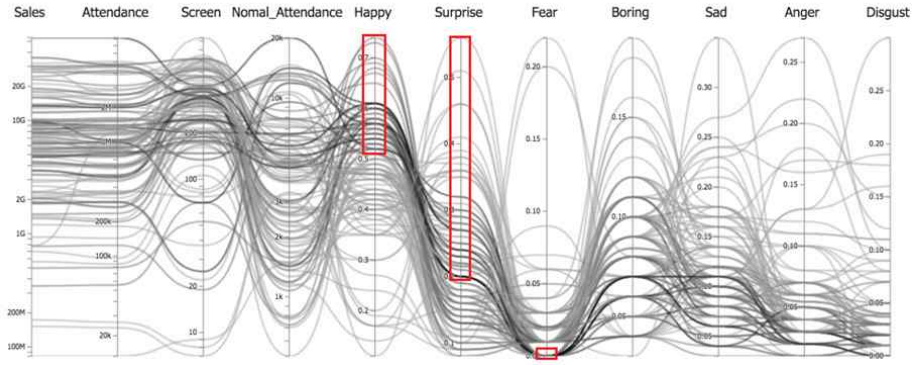

그림 47. 군집2에 대한 최종 Parallel coordinates

라. 군집 3 영화에 대한 의사결정나무분석 및 시각화 검증

군집3(Adventure, Meloromance, Action, Fantasy, Historicaldrama)에 속한 영화 데이터 집단에 대한 최적분리는 Happy에 의해 최조 이지 분리 되었다. 의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 9에 해당하는 집단의 분할 규칙은 Happy > 0.295 & Surprise > 0.275 & Sad > 0.145의 순서로 3번 분할 된 것을 확인 할 수 있다. 이번 장에서는 군집3 에서 최대의 흥행도 값이 예측된 노드 9에 대한 데이터 분할 과정을 Parallel coordinates시각화 분석 방법을 통해 검증하고자 한다. 노드 9에 대해서 분할 기준이 적용되지 않은 전체 값은 그림 48 과 같다.

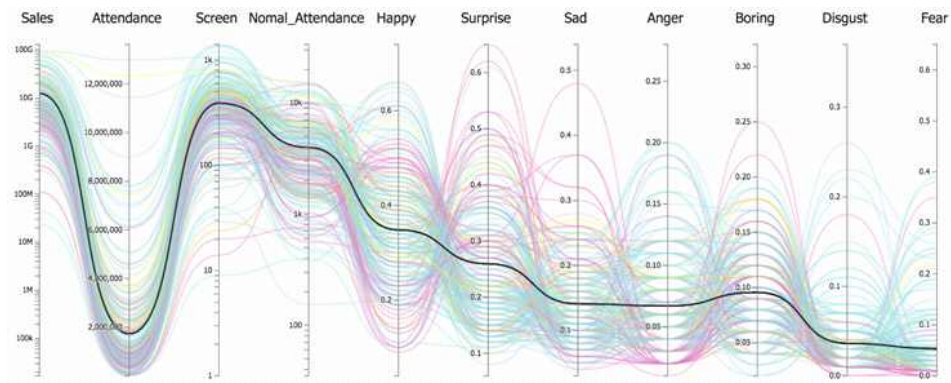

그림 48. 분할기준이 적용되기 전의 Parallel coordinates시각화

다음으로 Happy의 값이 0.295이상 일 때와 아닐 때로 최초 분할되었으며 Happy의 값이 0.295이상인 노드에 대한 시각화 결과는 그림 49 와 같다.

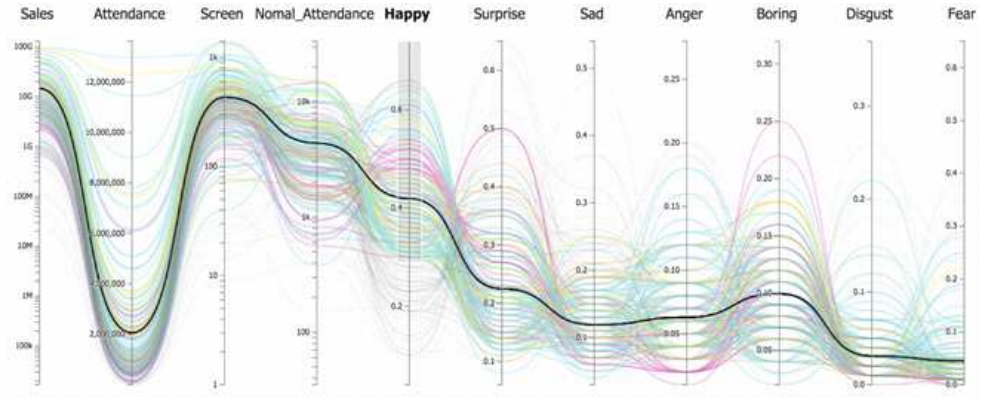

그림 49. Happy > 0.295이 적용된 시각화 결과

Happy를 기준으로 0.295이상인 데이터를 선택하였을 때 Happy의 값이 낮았던 멜로로맨스 장르의 영화가 51에서 27로 대폭 감소하였다. 두 번째로 적용된 분할 기준은 Surprise이며 Surprise값이 0.275이상 일 때와 아닐 때로 분류된 시각화 결과는 그림 50 과 같다.

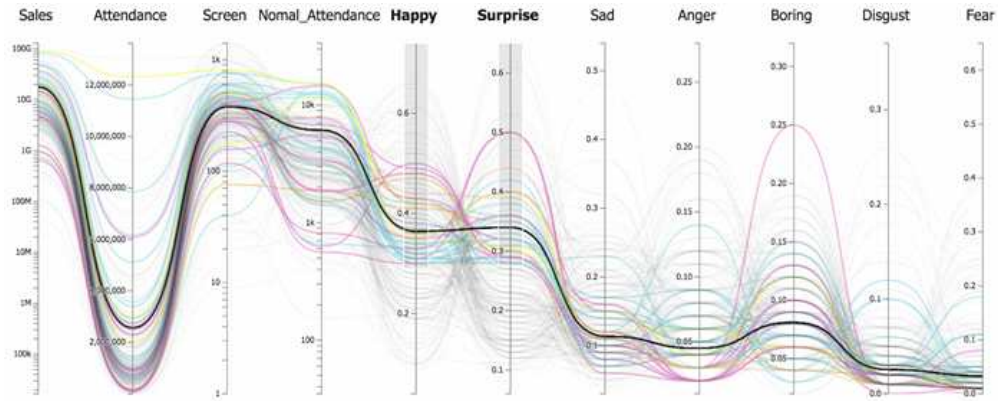

그림 50. Happy > 0.295 & Surprise > 0.275이 적용된 시각화 결과

Happy는 0.295이상이고 Surprise는 0.275이상인 데이터를 선택하였을 때 멜로로맨스와 액션 장르가 각각 11편, 18편으로 선택된 36편의 영화중 가장 높은 비율을 차지하였다. 마지막으로 적용된 분할 기준은 Sad이며 Sad값이 0.145이상 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각화 결과는 그림 51, 그림 52 와 같다.

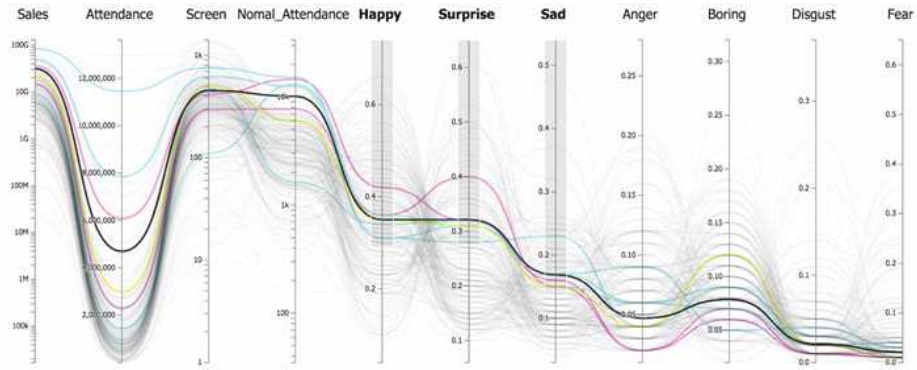

그림 51. 노드 9 에 대한 최종 Parallel coordinates

그림 52. 노드 9 에 최종 포함된 영화 정보

의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 9에 대해서 Parallel coordinates시각화 분석 방법을 사용하여 분석 한 결과, 노드 9 집단에 포함된 영화는 200 Pounds Beauty, All for Love, D-War, Fighter In the Wind, Haeundae, Hindsight, The Servant이었으며 영화의 장르는 Action이 4, Historicaldrama가 1, Meloromance가 2 포함된 것을 확인 할 수 있었다. 분할 과정에서 최종 까지 선택된 라인의 색상을 진하게 하여 표현한 전체 영화에 대한 시각화 모습은 그림 53 과 같다.

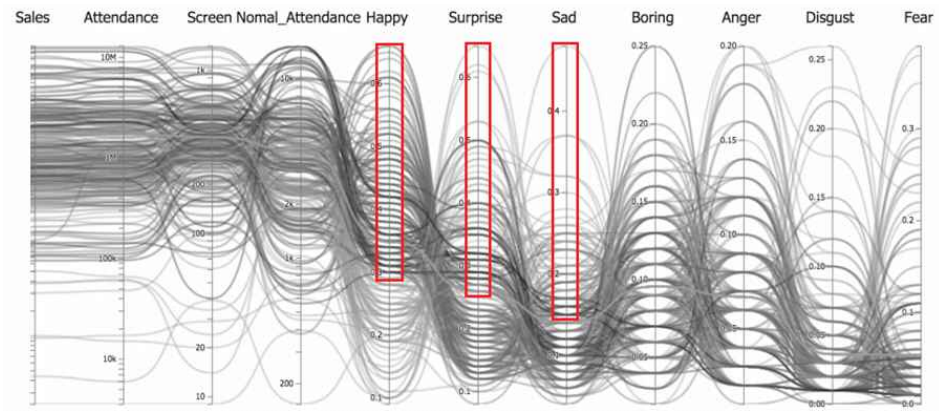

그림 53. 군집 3에 대한 최종 Parallel coordinates

마. 군집 4 영화에 대한 의사결정나무분석 및 시각화 검증

군집4(Drama, SF, Crime, War)에 속한 영화 데이터 집단에 대한 최적분리는 Boring에 의해 최조 이지 분리 되었다. 의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 9에 해당하는 집단의 분할 규칙은 Boring < 0.045 & Disgust < 0.035 & Fear > 0.015의 순서로 3번 분할 된 것을 확인할 수 있다. 이번 장에서는 군집4 에서 최대의 흥행도 값이 예측된 노드 9에 대한 데이터 분할 과정을 Parallel coordinates시각화 분석 방법을 통해 검증하고자 한다. 노드 9에 대해서 분할기준이 적용되지 않은 전체 값은 그림 54 와 같다.

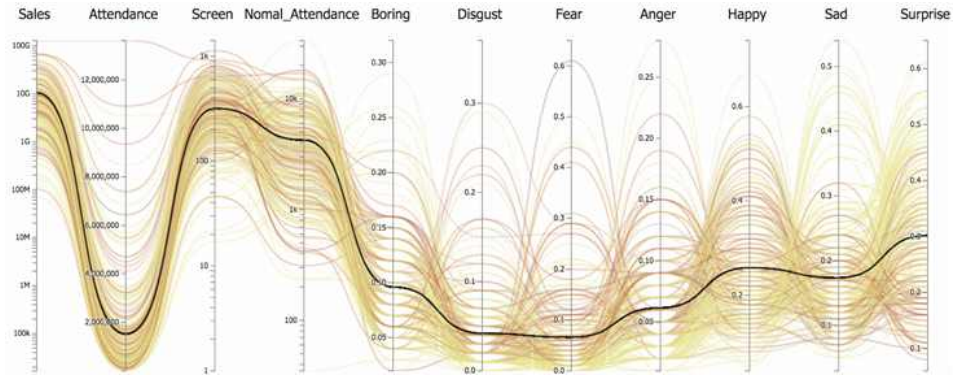

그림 54. 분할기준이 적용되기 전의 Parallel coordinates시각화

다음으로 Boring의 값이 0.045이하 일 때와 아닐 때로 최초 분할되었으며 Boring의 값이 0.045이하인 노드에 대한 시각화 결과는 그림 55 와 같다.

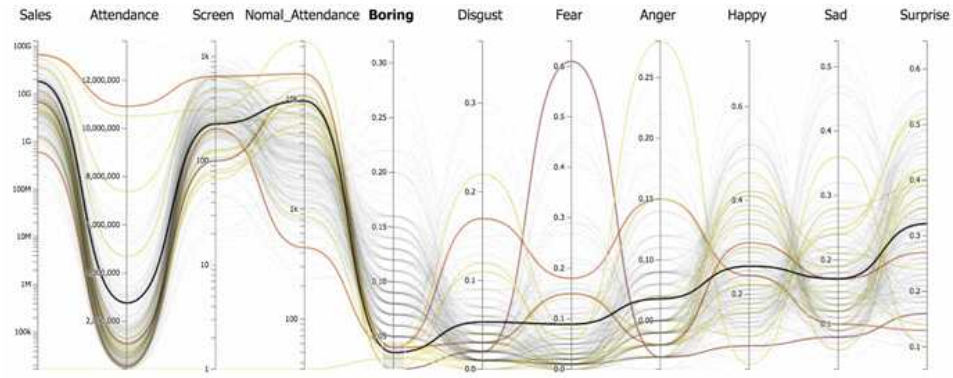

그림 55. Boring < 0.045이 적용된 시각화 결과

Boring를 기준으로 0.045이하인 데이터를 선택하였을 때 모든 장르의 영화가 대폭 감소하였고 범죄 장르의 영화는 완전히 제외 되었다. 두 번째로 적용된 분할 기준은 Disgust이며 Disgust값이 0.035이하 일 때와 아닐 때로 분류된 시각화 결과는 그림 56 과 같다.

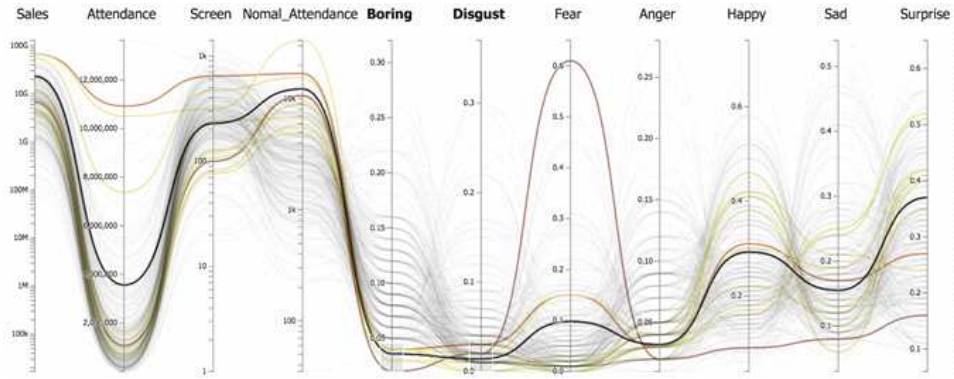

그림 56. Boring < 0.045 & Disgust < 0.035이 적용된 시각화 결과

Boring은 0.045이하이고 Disgust는 0.035이하인 데이터를 선택하였을 때 드라마 장르의 영화가 8편으로 선택된 10편의 영화중 가장 높은 비율을 차지하였다. 마지막으로 적용된 분할 기준은 Fear이며 Fear값이 0.015이상 일 때와 아닐 때를 추가적으로 적용하여 분류된 시각화 결과는 그림 57, 그림 58 과 같다.

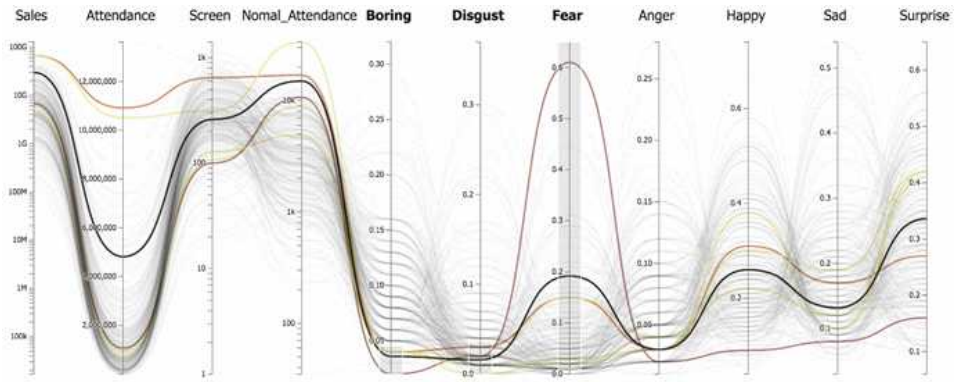

그림 57. 노드 9 에 대한 최종 Parallel coordinates

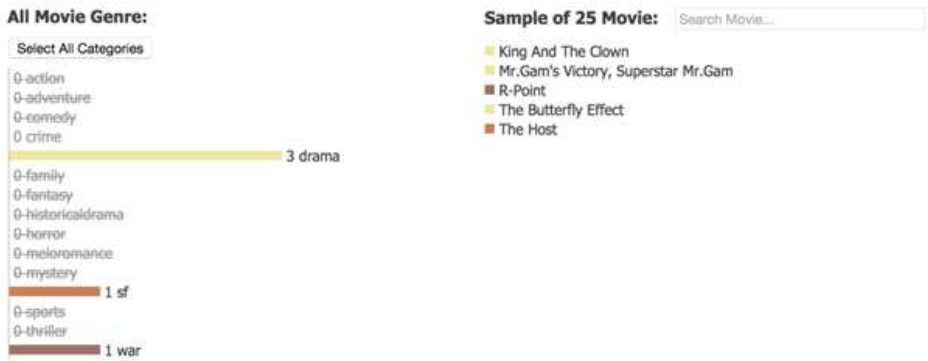

그림 58. 노드 9 에 최종 포함된 영화 정보

의사결정나무 분석을 통해 영화 흥행성의 예측 값이 높게 측정된 노드 9에 대해서 Parallel coordinates시각화 분석 방법을 사용하여 분석 한 결과, 노드 9 집단에 포함된 영화는 King And The Clown, Mr.Gam's Victory, Superstar Mr. Gam, R-point, The Butterfly Effect, The Host등 이었으며 영화의 장르는 Drama가 3, SF가 1, War이 1 포함된 것을 확인 할 수 있었다. 분할 과정에서 최종 까지 선택된 라인의 색상을 진하게 하여 표현한 전체 영화에 대한 시각화 모습은 그림 59 와 같다.

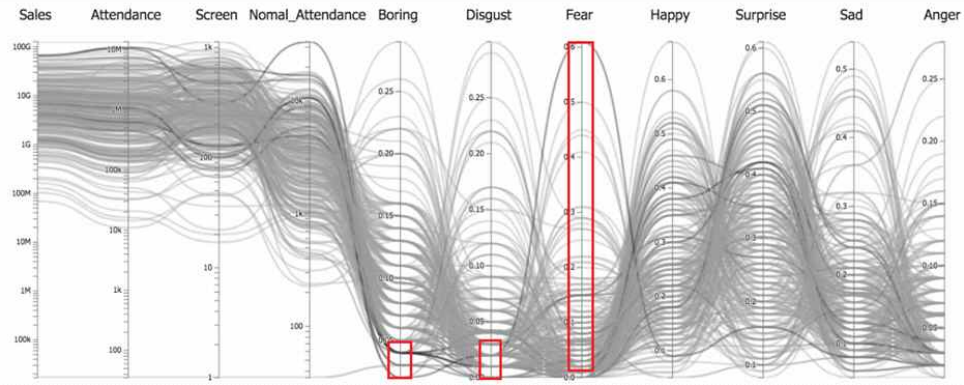

그림 59. 군집 4에 대한 최종 Parallel coordinates

바. 시각화 검증 결과에 대한 종합적인 해석

본 장에서는 선행된 의사결정나무분석을 통해 생성된 최종 예측 모형을 Parallel coordinates 시각화 방법을 통하여 검증하고 시각화 분석 방법을 결합하여 사용자가 유동적으로 분석 과정에 참여하는 방법을 제안하였다. 의사결정나무분석의 경우 패턴인식(Pattern recognition) 혹은 기계학습(Macine learning)으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 높은 신뢰성으로 목표가 되는 변수 값을 높게 예측할 수 있는 분할기준을 제시한다. 하지만 사용자의 입장에서 통계적인 지식이 부족할 경우 분석된 결과 외에는 알 수 없다는 한계가 있다.

본 연구에서는 이러한 한계를 보안하기 위하여 의사결정나무분석을 통해 제시된 분할기준을 Parallel coordinates를 활용하여 사용자가 직접 분류하는 방법을 제시하였다. 분석과정에 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. 군집3(Adventure, Meloromance, Action, Fantasy, Historicaldrama)의 경우

첫번째 분할 기준으로 Happy의 값이 0.295이상인 노드가 선택 되었을 때 Happy의 값이 낮았던 멜로로맨스 장르의 영화가 51에서 27로 대폭 감소한 결과와 두번째 분할 기준으로 Surprise값이 0.275이상인 데이터를 선택하였을 때 Surprise값이 낮았던 액션 장르의 영화가 88에서 18로 대폭 감소하는 것을 확인 할 수 있었다. 이는 군집된 집단 내에서도 장르에 따라서 감정어휘 값들의 비중이 서로 상이하고 해석 될 수 있다. 또한 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. 군집1(Horror, Mystery, Thriller)에 대한 의사결정나무분석 결과에서 가장 높은 흥행도 예측 값을 보인 노드 16에 포함되는 데이터들의 특성을 확인하여 본 결과는 다음과 같다. 그림 60 과 그림 61 은 노드 16에 포함된 영화중에서도 흥행도 값이 높았던 영화 The Big Swindle과 집단에 포함된 영화 내에서 흥행도 값이 낮았던 영화 Fast Five를 나타내고 있다. 두 영화의 감정어휘 분포는 서로 비슷하지만 Anger의 경우 흥행도가 높았던 영화는 Anger값이 낮았고, 흥행도가 낮았던 영화의 경우 Anger값이 높았던 것을 확인 할 수 있다.

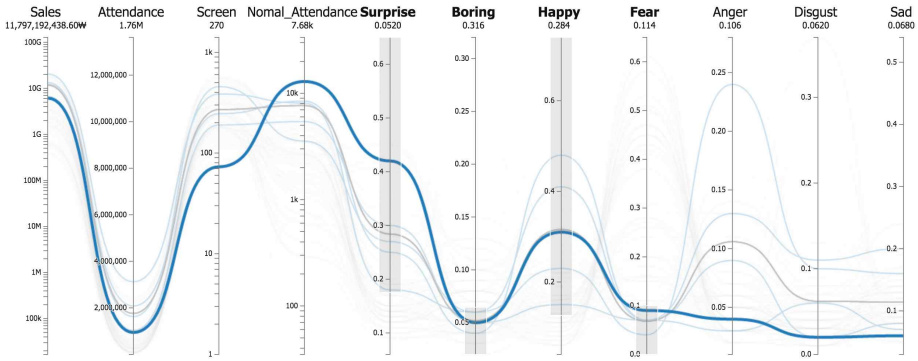

그림 60. 영화 The Big Swindle에 대한 시각화 결과

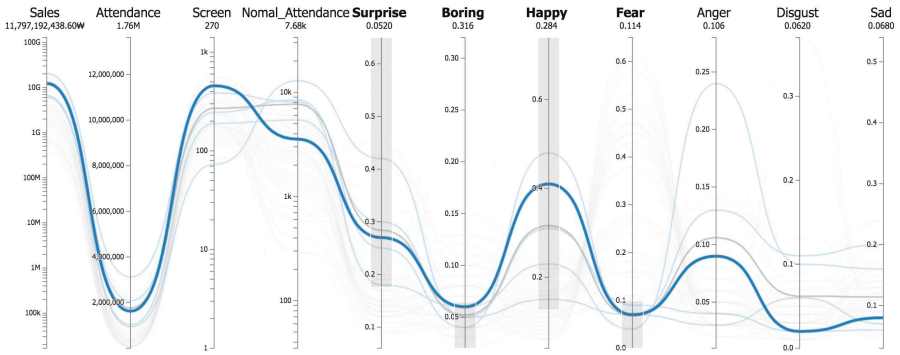

그림 61. 영화 Fast Five에 대한 시각화 결과

이처럼 통계적인 분석과 시각화 분석 방법을 결합하여 사용할 경우 각각의 분석 방법은 서로 보안 될 수 있으며 사용자는 데이터로부터 더욱 다양한 정보를 얻어 낼 수 있다.

Ⅵ. 결론

산업의 성장과 함께 방대한 양의 데이터들이 생산되었으며 생산된 데이터를 활용, 분석하여 가치 있는 정보를 추출하고, 현상을 예측하는 예측분석의 활용이 중요해지고 있다. 예측분석은 패턴인식(Pattern recognition) 혹은 기계학습(Macine learning)으로 불리는 확률적 학습 알고리즘을 기반으로 하기 때문에 분석 결과의 정확도와 신뢰성이 높다. 하지만 분석에 사용되는 알고리즘이 복잡하고 많은 조건을 가정해야하기 때문에 사용자가 분석과정에서 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 흥행성과 영화 리뷰에서 추출한 감정어휘 값으로 이루어진 데이터가 활용되었다. 영화의 흥행성을 예측하기 위해 예측분석의 한 종류인 의사결정나무분석을 수행하고 다양한 시각화 분석 기법 중에서 Parellel coordinates를 활용하여 예측모형을 검증하였다. Parellel coordinates는 데이터의 변수들을 각각의 축으로 설정하고 데이터마다의 특징을 하나의 라인으로 표현함으로써 개별 데이터의 특징을 파악하기 쉬우며 각각의 라인들이 겹쳐질 경우 데이터가 군집화 되는 패턴을 확인[36] 할 수 있기 때문에 예측 모형의 결과를 검증하기에 적합한 시각화 방법이라고 할 수 있다.

첫째, Parellel coordinates 시각화 분석을 활용하면 의사결정 나무 분석에서 제시된 예측모형의 분할 기준이 적용될 때 마다 변하는 데이터의 패턴을 파악할 수 있다. 예를 들어, 군집3(Adventure, Meloromance, Action, Fantasy, Historicaldrama)의 경우 첫 번째 분할 기준으로 Happy의 값이 0.295이상인 노드가 선택 되었을 때 Happy의 값이 낮았던 멜로로맨스 장르의 영화가 51에서 27로 대폭 감소한 결과를 확인하였다. 이는 Parellel 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. 또한 두 번째 분할 기준으로 Surprise값이 0.275이상인 데이터를 선택하였을 때 Surprise값이 낮았던 액션 장르의 영화가 88에서 18로 대폭 감소하는 것을 확인 할 수 있었다. 이도 위와 마찬가지로 Parellel coordinates의 기능을 활용하여 분석된 결과임을 알 수 있다. 이러한 결과는 일반적으로 군집된 집단 내에서도 장르에 따라서 감정어휘 값들의 비중이 서로 상이하다고 해석 될 수 있다. 또한 이를 본 연구에서 사용된 데이터가 아닌 일반적인 데이터에 빗대어 해석하면 데이터가 지니는 인구통계학적 특성

[36] Rick Walker.

에 따라서 데이터는 서로 상이한 특성을 지니고 있으며 적용되는 분할기준에 따라 선택, 제거되는 데이터의 특성도 변화한다고 할 수 있다.
둘째, 최종노드에 포함된 데이터들도 서로 상이한 특성을 지니고 있다는 것을 확인 할 수 있다. 군집1(Horror, Mystery, Thriller)에 대한 의사결정나무분석 결과에서 가장 높은 흥행도 예측 값을 보인 노드 16에 포함되는 데이터들의 특성을 확인하여 본 결과 노드 16에 포함된 영화중에서도 흥행도 값이 높았던 영화 The Big Swindle과 집단에 포함된 영화 내에서 흥행도 값이 낮았던 영화 Fast Five가 포함되어 있었다. 두 영화의 감정어휘 분포는 서로 비슷하지만 흥행도가 높았던 영화는 Surprise의 값이 높고 Anger값이 낮았으며, 흥행도가 낮았던 영화의 경우 Anger값이 높고 Surprise값이 낮았던 것을 확인 할 수 있다. 이러한 해석은 Parellel coordinates의 기능 중 개별 영화에 대한 검색과 선택의 기능을 활용한 분석 결과라고 할 수 있다. 이러한 결과를 통해 예측분석으로 도출된 최종 모형 내에서도 데이터 사이에 추가적인 관계가 존재한다는 것을 확인하였으며 시각화 분석을 사용할 경우 이러한 관계를 더 잘 확인 할 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 각 변수의 관계를 파악하고 높은 영화 흥행성을 예측하기 위한 예측모형을 도출하였으며, 시각화 분석에서는 변수들의 분포를 파악하는 사용자 인터렉션이 가능한 다양한 기능을 제공함으로서 최종적으로 제시된 예측모형을 검증하고 데이터로부터 더 다양한 정보를 도출하기 위한 방법론을 제시하였다.
향후 연구로써 단기적으로는 의사결정나무분석의 분할 순서에 따른 깊이(Depth)를 Parallel coordinate에서 나타낼 수 있는 방법에 대한 연구가 진행되어야 하며, 장기적으로는 본 연구에서 활용한 Parallel coordinate 방법뿐만 아니라 다양한 시각화 분석 방법을 통계 분석 방법과 결합함으로써 통계적 방법으로 도출하지 못한 데이터의 유의미한 의미를 파악하는 연구가 진행되어야 한다.

참고문헌

학위, 학술논문

▶ Adel Ahmed, Vladimir Batagelj, Xiaoyan Fu, Seok-Hee Hong, Damian Merrick, Andrej Mrvar, “Visualisation and Analysis of the Internet Movie Database”, Asia-Pacific Symposium on Visualisation 2007 IEEE, 2007.

▶ Litman, B, “Predicting Success of Theatrical Movies: An Empirical Study”, Journal of Popular Culture, 16 (Spring), pp.159-175, 1983.

▶ Hyoji Ha, Wonjoo Hwang, Hanmin Choi, Gi-nam Kim, Hansung Kang, Kyungwon Lee, “CosMovis: Semantic Network Visualization on Sentiment Words for Movie Recommendation System”, VIS 2014 Computer society IEEE, 2014.

▶ Adam Perer, Ben Shneiderman,"Integrating Statistics and Visualization: Case Studies of Gaining Clarity during Exploratory Data Analysis", CHI 2008 Proceedings · Visual Synthesis, pp. 265-274, 2008.

▶ Inselberg, A, "The plane with Parallel coordinates", The Visual Computer, pp. 69-91, 1985.

▶ Parrott, W, "Emotions in Social Psychology", Philadelphia: Psychology Press, 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, pp. 667-672, 2003.

▶ Borg, I., Groenen, P., "Modern Multidimensional Scaling: theory and applications", New York: Springer-Verlag, pp. 207-212, 2005.

▶ Lijun Yin, Xiaozhou Wei, "Multi-Scale Primal Feature Based Facial Expression Modeling and Identification", Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp.1-6, 2006.

▶ Challaglla, Goutam N. & Shervani, Tasaddug A., Dimensions and Types of Supervisory Control: Effectson Salesperson Performance and Satisfaction, Journal of Marketing 60(1), pp.89-105, 1996

▶ E. Kandogan, "Visualizing Multi-Dimensional Clusters, Trends, and Outliers using Star Coordinates.", ACM SIGKDD ’01, pp. 107-116, 2001.

▶ 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, pp.36-44, 2013.

▶ Pak Chung Wong, J. Thomas, "Visual Analytics", IEEE Computer Graphics

and Applications Volume 24 Issue 5, pp. 20-21, 2004.
▶ David Lechevalier, Anantha Narayanan, Sudarsan Rachuri, "Towards a Domain-Specific Framework for Predictive Analytics in Manufacturing", 2014 IEEE International Conference on Big Data, pp. 987-995, 2014.
▶ DeGroot, Schervish, "Definition of a Statistic". Probability and Statistics Third Edition Addison Wesley, pp.370-371, 2002.
▶ 김연형, 홍정한, “영화 흥행 결정 요인과 흥행 성과 예측 연구”, 한국통계학회 논문집, 제18권, 제6호, pp.859-869, 2011.
▶ 박승현, 송현주, 정완규, “한국영화의 흥행성과 결정 요인에 관한 연구”, 언론과학연구, 제11권, 제4호, pp.231-258, 2011.
▶ 성영신, 박진영, 박은아, “온라인 구전 정보가 영화 관람 의도에 미치는 영향”, 광고연구, 제57권, pp.31~52, 2002.
▶ 김휴종, “한국영화스타의 스타파워분석”, 삼성경제연구소 연구보고서, 1997.
▶ 이준웅, 송현주, 나은경, 김현석, “정서 단어 분류를 통한 정서의 구성 차원 및 위계적 범주에 관한 연구”, 한국 언론 학보, 제52(1)권, pp.85-116, 2008.
▶ 김정호, 김명규, 차명훈, 인주호, 채수환, “한국어 특성을 고려한 감성 분류”, 감성 과학, 제13(3)권, pp.449-458, 2010.
▶ 하효지, 김기남, 이경원, “영화 리뷰의 감정 어휘 공간 및 영화 관람의 상황분석 연구”, 디자인 융복합 학회, 제12(6)권, pp.43-59, 2013.
▶ 한덕웅, 강혜자, "한국어 정서 용어들의 적절성과 경험 빈도", 한국 심리학회지, 제19권, pp.78-98, 2000.
▶ 권영란, 김세영, "의사결정나무분석 기법을 이용한 중학생 인터넷게임중독의 보호요인 예측", 정신간호학회지 13호, pp. 12-20, 2014.
▶ 최종후, 서두성, "의사결정나무를 이용한 개인휴대통신 해지자 분석", 한국경영과학회, pp. 377-380, 1998.
▶ 박지연, 전범수, “네티즌의 흥행 영화 리뷰에 포함된 감정 동사 이용 특성 연구”, 한국 콘텐츠 학회, 제14(5)권, pp.85-94, 2014.
▶ 박승현, 송현주, “영화의 흥행성과와 제작비 규모와의 관계 : 2011년 한국영화의 흥행 결정 요인 분석”, 사회 과학 연구 제51권 1호 , pp. 45~79, 2012.
▶ 정한두, “의사결정나무분석을 통한 중소형 아파트 거주세대의 이주와 리모델링 결정요인”, 대한 건축 학회, 제30권, pp.45-56, 2014.
▶ 손용정, 김현덕, “의사결정나무분석을 이용한 컨테이너 수출입 물동량 예측”, 한국 항문 경제 학회지, 제28권, pp.193-207, 2012.
▶ 성정연, 조광수, "비주얼 햅틱 형용사의 지각적 어휘 공간 연구", Design Convergence Study 38 Vol.12. no.1, pp.123-125, 2013.

관련 서적

▶ 경찰청, "지리정보 통합한 지리적 프로파일링 시스템 구축 (GeoPros)", 2013 빅데이터 사례집, pp.65-67, 2013.

웹사이트

▶ NAVER 영화, http://movie.naver.com
▶ 영화 진흥 위원회 , http://www.kobis.or.kr/kobis/business/mast/mvie/searchMovieList.do

Abstract

Recently, predictive analytics is becoming more important with the development of information and communication.

Predictive analytics is closely related our daily life.

However, 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, I could find some implications that haven't been found in the previous studies.

First, I could find data pattern when I adjust data selection according as splitting criteria for the decision tree method.

Second, I could find what type of data included in the final prediction model.

I could find 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 I verified final prediction model and suggested analysis method extract variety of information from the data.