논문 19-44-09-24 The Journal of Korean Institute of Communications and Information Sciences '19-09 Vol.44 No.09
https://doi.org/10.7840/kics.2019.44.9.1788

머신러닝 기반 자동 데이터 시각화를 위한 특징공학

최 희 원•, 김 한 준°

Meta-Feature Engineering for Machine Learning-Based Automated Data Visualization

Hee-won Choi•, Han-joon Kim°

요 약

본 논문은 머신러닝을 기반으로 한 자동 데이터 시각화 시스템의 실현을 목표로 삼고, 시각화 추천 모델을 구성하기 위한 메타 수준의 특징 공학 과정을 소개한다. 기본적으로 시각화 결과는 데이터 분석의 목적에 따라 달라질 수 있으며, 데이터에 대한 이해도가 커질수록 다양한 결과가 얻어질 수 있다. 이번 실험을 통해, 우리는 자동 시각화 시스템을 구축하기 위해 시각화 결과의 유의미성을 결정할 수 있는 다양한 메타특징 변수를 설계하고, 이를 사용한 시각화 추천 모델을 구성하였다. 성능 평가를 위해 'R datasets', 'UC Irvine Machine Learning Repository' 및 'Data.world'에서 제공하는 데이터셋을 사용하여, 의사결정나무 기반 자동 시각화 모델이 최상의 성능을 제공한다는 사실을 확인하였다.

Key Words : big data, data visualization, machine learning, feature engineering, meta data

ABSTRACT

This paper aims at realization of an automatic data visualization system based on machine learning, and introduces a metal-level feature engineering process to construct a visualization recommendation model. Basically, the visualization results can be varied according to the purpose of the data analysis, and as the understanding of the data becomes grow, more various results can be obtained. Through these experiments, we have designed various meta-feature variable to determine the significance of the visualization results in order to develop the automatic visualization system and constructed the visualization recommendation model using the meta-features. For performance evaluation, we have used three data sources including R datasets, UC Irvine ML Repository, and Data.world, and have found that the decision tree-based recommendation model provides the best performance.

Ⅰ. 서 론

요즘 빅데이터 시대의 도래와 더불어 데이터 시각화를 이용한 데이터 탐색 및 분석 방법은 필수적인 요소가 되었다. 데이터 시각화는 데이터를 단시간에 이해할 수 있는 효과적인 방법이며, 정보 전달에 있어서 큰 기여를 한다[1]. 특히 이는 데이터에 내재된 유용한 정보를 발견하고 상대방과 대화하는 강력한 수단이다. 사용자는 데이터에 대한 이해도가 높을수록 분석 목적에 따라 시각화의 유형을 변경하면서 다양한 결과

※ 본 연구는 국토교통부 도시건축연구 사업의 연구비지원(19AUDP-B100356-05)에 의해 수행되었으며, 또한 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2019-2018-08-01417)
• First Author : University of Seoul Department of Electrical and Computer Engineering, chw6221@uos.ac.kr, 학생회원
° Corresponding Author : University of Seoul Department of Electrical and Computer Engineering, khj@uos.ac.kr, 정회원
논문번호 : 201905-062-C-RE, Received April 30, 2019; Revised July 5, 2019; Accepted July 19, 2019

를 얻을 수 있다.

최근에는 데이터를 재구성하여 의미를 찾아내는 방법들이 제안되고 있다. 이는 올바른 의사결정을 할 수 있도록 지원하고 있으며, 가공되지 않은 원천 데이터로부터 의미 있는 정보를 직접적으로 얻어내기 위함이다[1].시각화 시스템이 정확한 정보 전달을 수행하기 위해서는 체계적인 시각화 설계 과정이 요구된다[2]. 현재까지 다양한 방식의 시각화 기법이 제안되고 있는데, 이들은 대체로 다수의 데이터 탐색 단계를 가진다. 그 중에서 대표적 기법으로 Ben Fry는 그림 1.의 7단계 시각화 과정을 제안하였다[3].

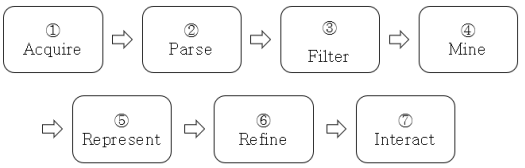

그림 1. Ben Fry의 7단계 시각화 과정
Fig. 1. Interactions between the 7 stages

① 획득(Acquire): 데이터에 대한 정보를 수집
② 분해(Parse): 사용자 의도에 맞게 데이터의 형식을 정리
③ 선별(Filter): 데이터의 의미있는 정보만 남기고 불필요한 정보 제거
④ 마이닝(Mine): 수학 및 통계적 기법을 활용하여 데이터 분석
⑤ 표현(Represent): 앞서 얻어진 정보를 통해 기본 시각화 모델을 선택
⑥ 정제(Refine): 선택한 기본 시각화 결과를 보완 및 개선
⑦ 상호작용(Interact): 사용자 데이터 간 상호 작용 기능을 추가

시각화에 익숙하지 않은 사용자는 Ben Fry가 제안한 시각화 방법론을 따르기가 수월하지 않다. 시각화 표현은 주어진 데이터에 대한 이해도가 높더라도 그 규모가 크다면 오랜 시간이 소요된다. 기존 시각화 서비스 도구는 사용자가 시스템에서 데이터의 특징 값과 차트를 선택한 후, 보여진 시각화의 좋고 나쁨을 판단한다. 이는 시각화 과정을 단축시키지만 사용자는 시각화 서비스 도구를 사용하기 위해 데이터에 대한 높은 이해도를 가져야한다. 대조적으로, 자동 데이터 시각화 시스템은 데이터에 대한 높은 이해도를 요구하지 않으며, 시각화에 익숙하지 않은 사용자에게 유의미한 시각화를 제안하는 것이 특징이다. 또한, 자동 시각화 시스템은 사용자에게 시각화에 대한 접근성과 표현의 용이성을 제공함으로써, 시간 단축과 비용 절감의 효과를 기대할 수 있다.

자동 데이터 시각화를 위해 시각화 결과에 대한 좋고 나쁨을 판단하는 것은 머신러닝을 통해 추천 모델을 생성함으로써 가능하다[4,5]. 본 논문은 머신러닝 기반 자동 시각화 시스템을 실현하기 위해서 시각화 분류(추천) 모델을 구성하는 유의미한 메타특징 변수에 대한 특징 공학 방안을 제시한다. 12가지 종류의 데이터에 대한 다양한 실험을 통해 의사결정나무(decision tree)를 포함한 다양한 머신러닝 알고리즘을 활용하여 시각화의 분류 성능을 비교하며 자동 시각화 시스템의 실현이 가능함을 보인다.

Ⅱ. 연구 배경

본 논문의 목적은 머신러닝 기반 자동 시각화를 위한 프로세스를 정의하고, 머신러닝에 사용되는 메타성 특징 데이터(meta data)를 설계하여 결과적으로 자동 시각화 시스템의 토대를 만드는 것이다. 자동 시각화 머신러닝 기반 시각화 시스템은 그림 2.와 같은 구조를 가진다. 이는 크게 2개 단계(phase)로 구성되며, 각 단계별 수행과정은 학습단계와 추천단계로 구분할 수 있다. 첫 번째, 학습단계(training phase)는 주어진 학습데이터로부터 메타성 특징 데이터(meta data)를 추출하여 분류(추천) 모델을 생성하는 작업을 포함한다. 두 번째, 추천단계(recommendation phase)는 앞서 학습단계에서 도출된 분류 모델에 기반하여 임의의 데이터에 적합한 시각화 결과를 추천하는 서비스를 포함한다.

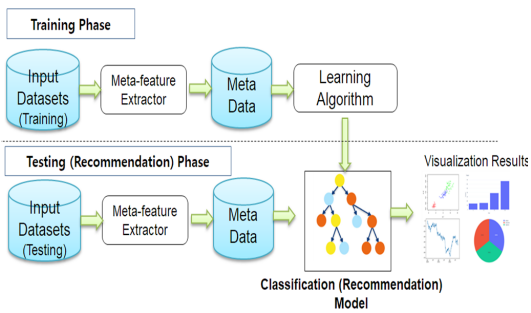

그림 2. 자동 시각화 시스템 아키텍처
Fig. 2. Overall architecture of automated visualization system

① 학습단계
- 입력데이터는 메타특징 추출기를 통해 메타데이터 데이터베이스를 구성한다. 그리고 메타데이터는 6가지 머신러닝 알고리즘을 활용하여 가장 적합한 시각화를 보여주는 추천 모델을 생성하고, 이 중에서 최선의 모델을 선택한다.

② 추천단계
- 임의의 데이터는 학습단계에서 추출한 메타특징과 같은 방법으로 추출기를 통해 메타데이터가 생성된다. 추출한 데이터는 학습단계에서 선택한 모델에 입력된 후, 합당한 시각화 결과를 추천해 준다.

본 연구는 자동 데이터 시각화 프로세스를 2개 단계(phase)로 구분하며, 정확한 시각화 추천 모델을 생성하기 위해 입력데이터로부터 메타특징 변수에 해당하는 양질의 메타데이터를 구성하는 것이 핵심이다.

Ⅲ. 특징 공학 및 시각화 기법

3.1 입력데이터

본 논문은 세 곳의 출처로부터 데이터를 제공받는다. R의 내장 데이터(R Datasets), UCI ML 저장소(UC Irvine Machine Learning Repository), Data.world에서 제공하는 총 12개의 데이터셋을 입력데이터로 사용한다.

표 1.은 입력데이터의 출처와 세부 데이터의 명칭을 보여준다. 입력데이터는 결측치 값이 없는 정형데이터로 준비한다. 컬럼의 개수는 2개에서 15개에 이르며, 레코드의 개수는 최소 32개에서 최대 17,585개에 이른다. 입력데이터의 컬럼 정보는 메타데이터 구성 요소로 중요한 역할을 한다. 이는 3.2절에서 메타특징 공학의 정의를 다루며 자세히 살펴본다.

표 1. 12개 입력데이터의 출처와 이름
Table 1. Source and name of 12 input data

Source	Name
R Datasets	Iris
	Airquality
	BostonHousing
	Orange
	Chick
	Mtcars
	USArrests
	Titanic
	Swiss
UC Irvine ML Repository	Airquality
	Wine
Data.world	Ebola

3.2 자동 시각화를 위한 메타특징 공학

특징 공학의 목적은 주어진 입력데이터의 도메인 지식을 활용하여 새로운 특징 값을 가공하는 것이다. 본 논문에서 구성한 시각화 메타특징은 시각화 대상이 되는 데이터로부터 시각화에 도움이 되며, 인간의 선호도 및 합당성을 학습하기에 적절한 메타데이터 변수를 의미한다. 이는 주어진 데이터의 기초 통계량, 분포, 컬럼 타입 등의 정보를 포함하며, 이러한 메타데이터의 추출 작업은 메타특징 공학(meta-feature engineering)이라 부른다. 따라서, 메타특징 공학은 대상으로부터 필요한 정보를 추출하는 작업이 중요하다.

유의미한 시각화를 포착할 수 있는 방법은 인간의 인식이다. 컴퓨터는 인간의 인식을 그대로 받아들이지 못하기 때문에, 이를 해석할 수 있도록 데이터화 시키는 작업이 필요하다. 시각화 메타특징 공학은 시각화 추천에 앞서 해당 시각화의 좋고 나쁨을 판단할 수 있는 특징 값을 입력데이터로부터 추출하는 과정을 거친다. 이는 사용자가 판단한 좋은 시각화의 패턴을 학습하기 위함이다. 시각화 대상이 되는 입력데이터는 적어도 한 가지 이상의 유의미한 시각화로 표현할 수 있다. 우리는 적합한 시각화 표현에 필요한 특징 값이 무엇인지 고찰하였으며, 이를 메타특징으로 구성한다. 그리고 자동 시각화를 위해 머신러닝 알고리즘을 활용할 때, 생성되는 학습모델을 구성하는 인자인 메타특징의 품질이 시각화 추천의 정확도 성능을 결정하게 된다.

본 연구에서 도출된 메타특징은 입력데이터셋으로부터 도출된 26개의 독립 변수와 이진 클래스를 갖는 종속 변수를 포함한다. 이것의 자세한 내용은 표 2.에 제시되었다. 우리가 설계한 메타특징은 DeepEye[4]에서 제안한 메타특징 14개에 새로운 13개의 특징을 추가한 것이다. 우리가 설계한 메타특징 변수에 기반한 추천 모델이 DeepEye의 추천 모델보다 우수함을 실험을 통해 입증할 것이다.

표 2.의 메타특징 중 1번, 2번은 2차원 시각화를 표현하는 가로축, 세로축 정보를 나타낸다. 2차원 시각화 가로축인 x축은 x_1으로, 세로축인 y축은 x_2로 정리함으로써, 독립 변수와 종속 변수의 표현과 혼동하지 않는다. 3번부터 19번은 통계량의 정보와 변수간의 상관계수 정보를 입력한다. 통계량 정보는 최소값,

표 2. 자동 시각화 모델을 위한 메타특징
Table 2. Meta-feature for Automated Visualization Model

No	Meta-feature	
1	Independent variable	Name of x-axis(x_1)
2		Name of y-axis(x_2)
3		Min(x_1)
4		Min(x_2)
5		Max(x_1)
6		Max(x_2)
7		Range(x_1)
8		Range(x_2)
9		Mean(x_1)
10		Mean(x_2)
11		Median(x_1)
12		Median(x_2)
13		Skewness(x_1)
14		Skewness(x_2)
15		Kurtosis(x_1)
16		Kurtosis(x_2)
17		Standard Deviation(x_1)
18		Standard Deviation(x_2)
19		Correlation(x_1, x_2)
20		Number of columns
21		Number of records
22		#Columns/#Records
23		Categorical(0,1)
24		Numerical(0,1)
25		Series(0,1)
26		Visualization Type (0,1,2,3)
27	Dependent variable	Bad=0, Good=1

최대값, 범위값, 평균값, 중간값, 왜도, 첨도, 표준편차를 나타낸다. 20번, 21번은 입력데이터의 컬럼과 레코드의 개수를 입력하며, 22번은 입력데이터의 컬럼을 레코드로 나눈 값으로 구성 비율(#Columns/#Records)을 나타낸다. 23번, 24번, 25번은 입력데이터에서 추출한 인자의 속성 정보를 나타낸다. 이는 범주형, 수치형, 시계열로 나뉘며, 해당 값은 포함유무에 따라서 미포함은 0, 포함은 1로 매핑 된다. 시각화 종류는 (0, 1, 2, 3)으로 산점도는 0, 막대그래프는 1, 라인그래프는 2, 파이그래프는 3으로 매핑 된다. 종속 변수는 입력데이터를 시각화 했을 때, 시각화 전문가가 좋고 나쁨을 판단하여 이를 1, 0의 값으로 직접 입력한다. 예를 들어, 그림 3.은 입력데이터셋 중 Titanic 데이터에서 임의로 입력데이터 특징을 선택하여 비교한 파이그래프이다. 인간의 인식은 파이그래프의 좋은 예(그림 3.(a) 참조)와 나쁜 예(그림 3.(b) 참조)로 판단할 수 있다. 이는 특정변수로 비교해볼 수 있으며, 좋은 예(그림 3.(a) 참조)의 경우 범주형 특징으로 좌석의 등급별 빈도수를 보여주고, 나쁜 예(그림 3.(b) 참조)의 경우 연속형 속성을 갖고 있기 때문에 Stephen few의 연구[8]에서 제안하는 좋은 시각화 조건에 적합하지 않다.

이처럼 시각화의 좋은 예를 판단하기 위해서는 다양한 입력데이터로부터 특징 값의 속성을 파악하고, 해당 속성 인자의 성분이 무엇인지 학습한다. 이를 자동 시각화 모델을 위한 메타특징의 성분으로 구성함으로써, 좋은 시각화 표현의 유의미한 패턴을 학습할 수 있어야한다.

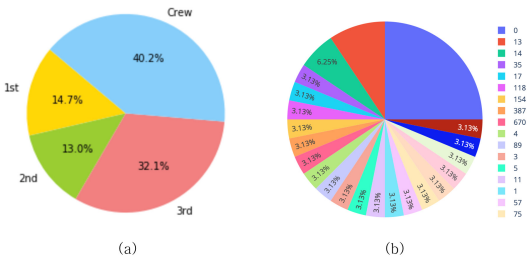

그림 3. 좋은 시각화의 예(a)와 나쁜 시각화의 예(b)
Fig. 3. Example of 'Good(a)' and 'Bad(b)' visualization

3.3 시각화 종류 및 표현 기법

데이터 시각화는 데이터의 특징 값을 적절한 시각화 요소로 표현하여 정보를 전달하는 것이다[3]. 이는 사용자에게 효과적인 메시지를 전달하기 앞서 표현 목적을 명료하게 구분지어야 한다. 본 연구는 표 3.에서 제시한 바와 같이 많이 사용되는 4가지 유형의 시각화인 산점도, 막대그래프, 라인그래프, 파이그래프로 정리한다. 이처럼 본 논문에서 사용한 시각화는 쓰임 목적에 따라 4가지 종류로 구분할 수 있다. 이는 Andy Kirk의 시각화 방법론[3], Muzammil Khan의 연구[6], Adobe[7]에서 서술한 내용을 종합하여 표 3.에 제시한다.

표 3. 목적에 따른 시각화의 종류
Table 3. Types of visualization according to purpose

No	Purpose	Type
1	Comparison	Bar chart Line chart
2	Relationship	Scatter plot
3	Distribution	Scatter plot
4	Composition	Pie chart

① 산점도(Scatter plot): 2개의 수치형 변수데이터를 2차원 공간에 표현하여 두 변수의 함수 관계를 예상하거나 데이터의 분포를 확인한다.
② 막대그래프(Bar chart): 범주형 특징 값을 포함하고, 주어진 값들이 뚜렷한 차이를 보이는 경우 혹은 상대적인 차이를 한눈에 알아본다.
③ 라인그래프(Line chart): 시계열 특징 값을 갖는 데이터를 시각화 할 때 용이하다.
④ 파이그래프(Pie chart): 전체 특징 값 중 범주형 특징 값이 차지하는 비율을 확인할 때 사용한다.

입력데이터 인자의 속성 정보는 메타특징 변수를 구성하고, 시각화를 구분할 수 있는 특징이다. 그림 4.는 시각화 기법 흐름도를 나타낸다. 특징 변수 속성 정보의 포함 유무는 그림 4.와 같이 시각화의 종류 4가지로 구분할 수 있으며, 이는 표 2.에서 언급한 메타특징 23번, 24번, 25번의 포함유무로 나타낸다.

첫 번째 단계는 수치형 변수를 포함할 경우 두 번째 단계로 넘어가며, 포함하지 않는 경우는 파이그래프를 내보낸다. 두 번째 단계는 수치형 변수를 포함하고, 범주형 변수의 포함유무에 따라 시각화의 유형을 결정한다. 수치형 변수와 범주형 변수를 모두 포함할 경우 막대그래프, 라인그래프로 나타나며, 포함하지 않을 경우 산점도로 내보낸다. 마지막 단계는 시계열 변수를 포함할 경우 라인그래프, 포함하지 않을 경우 막대그래프로 내보낸다. 따라서, 입력데이터 인자의 속성 정보를 담는 메타특징은 시각화의 패턴을 학습할 수 있는 일부 인자로 작용한다.

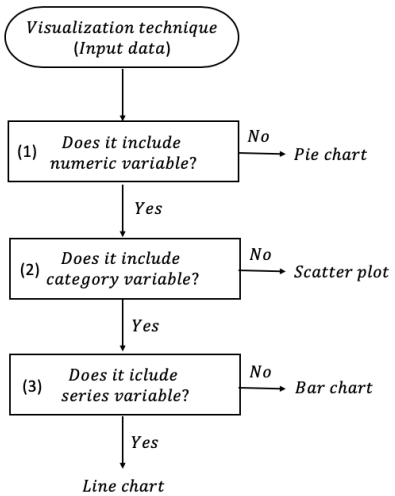

그림 4. 시각화 축의 속성 정보에 따른 시각화 흐름도
Fig. 4. Visualization flowchart based on information of axis

3.4 시각화 종류별 특징변수의 고찰

본 연구에서 우리는 다양한 실험을 통해 시각화에 대한 인간의 인식을 정량적 데이터로 표현할 수 있는 메타특징을 도출하였다. 시각화 유형은 표 3.과 같이 사용 목적에 따라 크게 4가지로 나뉠 수 있으며, 메타특징으로 이를 구체화 할 수 있다. 시각화는 표현 방법이 다양하기 때문에 특징변수에 관한 연구들이 많이 소개되어 있으며[2,6-8], 일반 사용자가 시각화 유형을 결정하기 위해서는 특징변수에 대한 정보를 숙지하고 있어야 한다. 다음은 입력데이터의 메타특징이 4가지 시각화 유형에 미치는 영향과 요구 조건을 산점도, 막대그래프, 라인그래프, 파이그래프 순으로 설명한다[6,7].

산점도는 두 개의 특징변수 사이의 분포와 관계를 확인할 때 사용한다. 분포는 데이터의 위치를 알 수 있으며, 관계는 상관계수로 그 의미를 파악할 수 있다[6,9]. 특징변수 사이의 관계는 상관계수 값으로만 상관성을 판단할 수 없기 때문에 시각화로 확인하는 과정이 필요하다.

산점도의 가로축과 세로축은 수치형 속성 인자를 갖는다. 그림 5.는 입력데이터셋 중 Mtcars 데이터의 상관계수를 이용한 산점도 시각화의 좋은 예(그림 5.(a) 참조)와 나쁜 예(그림 5.(b) 참조)를 나타낸다. 그림 5.(a)의 경우, 상관계수 -0.71을 갖고, 그림 5.(b)의 경우, 상관계수 0.74를 갖는다. 이를 산점도로 비교하였을 때, 전자는 후자보다 더 낮은 상관계수를 갖지만 유의미한 시각화를 표현한다. 이를 통해, 산점도는 상관계수로만 좋은 산점도의 경우를 판단하는 지표가 될 수 없기 때문에 표현된 시각화마다 전문가의 확인이 필요하다.

결과적으로 산점도는 상관계수로만 표현된 시각화의 좋고 나쁨을 판단하는 기준이 될 수 없으며, 상관계수를 제외한 다른 특징 값들의 유의미한 패턴을 통해 학습할 수 있도록 하는 것이 중요하다.

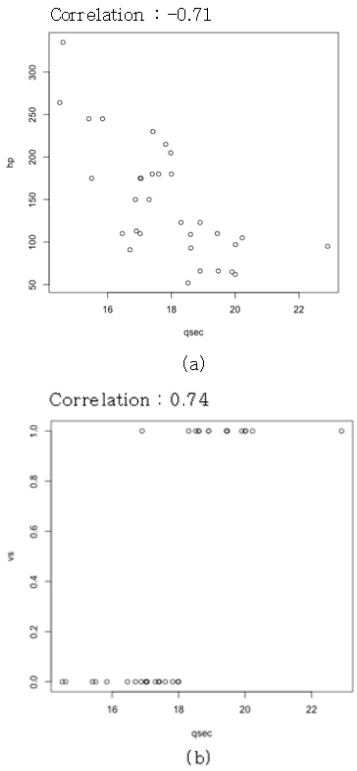

그림 5. 산점도 시각화 표현의 좋은 예(a)와 나쁜 예(b)
Fig. 5. Example of 'good' and 'bad' visualization using correlation coefficient

막대그래프는 흔히 그룹화 된 정보를 비교할 때 사용한다. 가로축은 범주형, 시계열형을 갖고, 세로축은 수치형이어야 한다[6,7]. 예를 들어, 입력데이터의 특징 중에서 범주형 또는 시계열 특징을 선택할 때, 수치형 컬럼의 최소값, 최대값, 평균값, 중간값은 막대그래프를 그리는데 적정한 기준을 제공한다.

그림 6.은 입력데이터셋 중 housing 데이터를 막대그래프로 표현한 좋은 예(그림 6.(a) 참조)와 나쁜 예(그림6.(b) 참조) 시각화를 보여준다. 입력데이터의 특징 변수는 수치형, 범주형 포함 유무에 따라 좋고, 나쁜 시각화인지 결정 인자로 작용하지만 해당 특징 변수의 포함 유무만으로 판단할 수 없기 때문에 전문가의 확인이 필요하다.

그림 6.(a)는 x축을 유일값이 5개인 범주형 특징을 입력하였고, y축은 빈도를 나타내는 수치형 특징을 입력하였다. 그림 6.(b)는 x축을 범주형 특징으로 입력하였으며 유일값의 개수가 5926개로 나타난다. 그림 6.(a)는 비교적 적은 유일값으로 표현한 막대그래프를 통해 차이를 한눈에 구분할 수 있다. 하지만 그

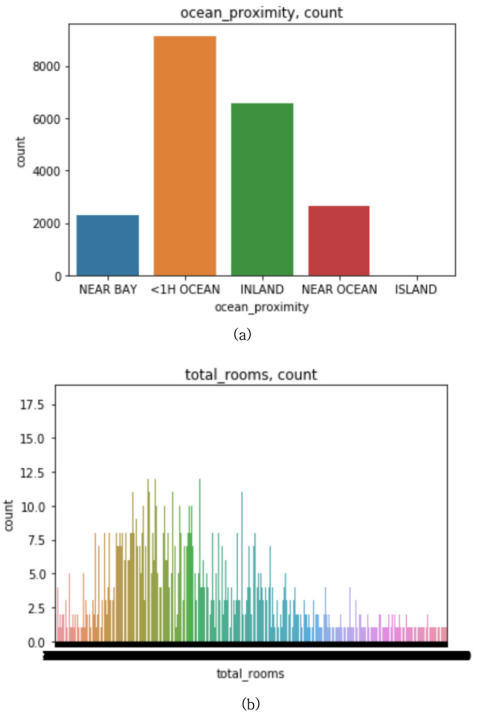

그림 6. 막대그래프 시각화 표현의 좋은 예(a)와 나쁜 예(b)
Fig. 6. Example of 'good' and 'bad' visualization of Bar chart

림 6.(b)는 5926개의 유일값을 x축에 표현함으로써 차이를 명확히 구분하기 어렵다. 이는 막대그래프의 경우 수치형과 범주형 특징이 포함 정보 외에도 범주형 특징의 유일값에 따라 좋고, 나쁜 시각화를 분류할 수 있다. 따라서, 막대그래프는 앞서 언급한 통계량 및 특징변수 외에도 숨겨진 유의미한 패턴의 특징 값을 머신러닝 알고리즘을 이용해 학습한다.

라인그래프는 과거에서 현재 또는 특정 기간까지의 추세 패턴을 찾아낸 것을 통해 미래를 예측하거나 변화를 해석할 때 사용 한다[6,7]. 그래서 가로축은 연도, 날짜, 시간과 같은 시계열 속성을 담고, 세로축은 수치형 속성을 입력한다.

그림 7.은 입력데이터셋 중 Airquality 데이터를 라인그래프로 표현한 좋은 예(그림 7.(a) 참조)와 나쁜 예(그림 7.(b) 참조) 시각화를 보여준다. 그림 7.(a)는 x축을 시계열 특징으로 입력하였고, y축은 수치형 특징으로 입력하여 라인 그래프를 표현하였다. 그림 7.(b)는 x축을 수치형 특징으로 바꾸고, y축은 그림 7.(a)와 동일한 수치형 특징으로 입력하였다. 그림 7.(a)는 시간이 흘러감에 따라 "Temp"의 변화를 해석

그림 7. 라인그래프 시각화 표현의 좋은 예(a)와 나쁜 예(b)
Fig. 7. Example of 'good' and 'bad' visualization of Line chart

할 수 있지만, 그림 7.(b)는 표현된 시각화에서 유의미한 정보를 얻어낼 수 없다. 라인그래프는 시계열 특징과 수치형 특징의 포함 유무에 따라 표현된 시각화가 좋고, 나쁨을 구분하는 기준을 제공한다. 따라서, 라인그래프는 사건이 순차적으로 발생하는 데이터의 패턴을 머신러닝 알고리즘을 이용해 학습한다.

파이그래프는 범주형 데이터가 차지하는 비율을 표현할 때 자주 사용하며, 바람직한 파이그래프가 되기 위해서는 선택된 특징에 대한 유일값의 개수가 적어야한다[6-8]. 만약에 하나의 특징 값이 포함하는 유일값의 개수가 많다면 그림 3.(b)에서 보는 바와 같이 나쁜 시각화로 표현된다.

그림 8.은 입력데이터셋 중 iris 데이터의 특징값의 하나인 "species"를 파이그래프로 보여준다. 범주형 특징을 갖는 "species"는 유일값이 3개이며, 그림 8.과 같이 일정 비율로 나누어진다. 이는 특징값의 구성 차이가 없다는 것을 확인할 수 있다.

파이그래프는 특징을 구성하는 변수들의 비율 차이를 확인할 수 있으며, 이를 통해 유의미한 결과를 얻어낸다. 따라서, 파이그래프 추천 모델은 입력데이터의 컬럼 수, 레코드 수, 상대적 비율 등을 활용하여 적합한 시각화 정보가 무엇인지 학습하는 것이 바람직하다.

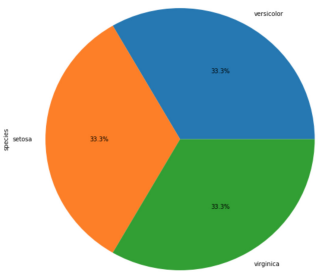

그림 8. 파이그래프의 예
Fig. 8. Example of piechart

Ⅳ. 메타특징을 이용한 머신러닝 기반의 시각화 추천

4.1 실험 환경

본 실험에서 사용한 입력데이터는 표 1.에서 언급한 12개 데이터셋이다. 앞서 3.2절에서 제안한 방법을 기반으로 입력데이터로부터 추출한 메타특징 27개와 2460개 레코드를 정제 및 처리하는 과정을 서술한다. 또한, Scikit-learn[10], Keras[11]에서 제공하는 라이브러리를 이용하여 적합한 분류(추천) 머신러닝 알고리즘을 찾고, 이중 성능이 가장 좋은 모델을 선택한다. 그림 9.는 주어진 입력데이터로부터 27개 메타특징에 해당하는 데이터를 추출하고, 이를 정제하여 학습에 필요한 학습데이터를 구축하여 시각화 종류별 분류 과정을 나타낸다.

① 입력데이터셋으로부터 27가지 메타특징에 해당하는 데이터가 산출된다.
② 산출한 메타특징 값의 범위를 Min-Max scaling을

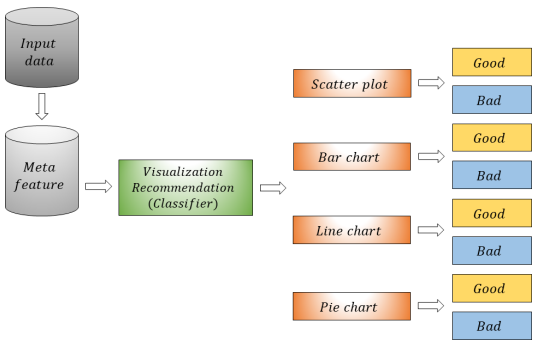

그림 9. 시각화 종류에 따른 분류 과정
Fig. 9. Categorization process by visualization type

사용하여 조정한 후, 시각화 유형별로 색인한다.

③ 시각화 유형별로 레코드 개수를 확인한 후, 클래스 'Good'과 'Bad'의 비율을 조정한다. 예를들어, 산점도의 경우, 레코드 606개에 클래스 'Good'과 'Bad'의 개수가 각각 91개, 515개로 나타나며, 이러한 클래스 불균형(imbalance)상태를 해소해야 한다. 이를 위해 'Bad'에 해당하는 레코드를 무작위로 제거함으로써 클래스 'Good'과 'Bad' 비율을 대략 1:1.5로 맞춘다.

④ 27개 메타특징으로 구성된 데이터는 학습데이터와 테스트데이터로 분할되며, 그것의 비율은 0.75:0.25이다.

⑤ 시각화 유형별로 클래스 분류 성능이 가장 높은 4가지 모델을 선택한다.

4.2 성능 평가

본 실험은 표 2.에서 언급한 메타특징 데이터를 기반으로 6가지 알고리즘을 사용하여 이진 분류 실험을 수행한다. 이진 분류는 주어진 분류 규칙에 따라서 표현된 시각화를 두 그룹으로 나눈다. 우리는 시각화 전문가가 표현된 시각화에 대해서 좋고 나쁨으로 분류한 메타특징 데이터를 기반으로 성능을 구성한다. 실험에서 사용한 메타특징 데이터는 학습데이터, 테스트데이터로 분리하고, 테스트셋 615개로 분류 성능을 확인한다. 테스트셋은 랜덤으로 추출하였으며, 시각화 전문가가 분류한 정보를 기반으로 성능을 확인한다. 해당 실험은 시각화 종류별로 4가지 분류(추천) 모델을 생성하기 때문에, 최선의 분류 성능을 가진 알고리즘을 선택한다. 성능 평가 지표는 $F_1-score$를 사용하며, 이는 정밀도(precision)와 재현율(recall)의 조화평균을 나타낸다. 정밀도(precision)는 정답으로 예측한 내용 중에 실제 정답으로 나타난 비율을 뜻하고, 재현율(recall)은 모델이 얼마나 정확하게 정답을 찾았는지 나타낸다. 기존 DeepEye에서 사용한 알고리즘은 3가지로 진행하였으며, 우리는 6가지 알고리즘으로 확장하여 실험한다. 6가지 알고리즘의 종류는 Decision Tree(DT), Random Forest(RF), Deep Neural Network(DNN), Logistic Regresion(LR), Support Vector Machine(SVM), Naive Bayes(NB)이다.

표 4.는 시각화 종류에 따른 6가지 머신러닝 학습 알고리즘의 성능을 나타낸다. S는 산점도, B는 막대그래프, L은 라인그래프, P는 파이그래프를 의미한다. 시각화는 3.4절에서 서술한 내용과 같이 종류별로 논의 된 규칙을 따라야 한다. 따라서, 6가지 알고리즘 중에서Decision Tree(DT)의 $F_1-score$가 가장 높게 나온 것을 확인 할 수 있고, 이는 규칙 기반의 분류 알고리즘이기 때문에 다른 5가지 알고리즘보다 더 높은 성능의 결과값을 가진다는 것을 알 수 있다[4].

그림 10.은 시각화 종류별 6가지 머신러닝 알고리즘 중에서 상위 3개(Top-3) 알고리즘의 성능을 비교한다. 이는 DNN(deep neural network), RF(random forest), DT(decision tree) 순으로 분류(추천) 성능이 증가하는 것을 다시 한번 확인할 수 있다. 산점도(S)는 6가지 학습 알고리즘 중에서 DT(decision tree)의 성능이 94.5%로 가장 높게 나왔으며, 막대그래프(B), 라인그래프(L), 파이그래프(P)의 경우도 94.3%, 100.0%, 94.7%의 값으로 DT(decision tree)의 성능이 가장 높은 결과로 나왔다. 따라서, 실험 결과 시각화 종류별 가장 높은 분류(추천) 성능을 보인 모델은 Decision Tree(DT)로 나타났다.

표 5.는 Decision Tree(DT)의 시각화 종류별 분류 성능을 확인 할 수 있는 성능 평가 지표이다. 이는 정

표 4. 시각화 종류별 6가지 머신러닝 알고리즘의 성능 비교
Table 4. Performance comparison of classifying visualization types with 6 machine learning algorithms

	F_1-score (%)			
	S	B	L	P
DT	94.5	94.3	100.0	94.7
RF	90.7	92.5	100.0	90.5
DNN	80.3	92.5	100.0	85.6
LR	75.4	70.5	100.0	80.2
SVM	74.1	75.2	100.0	82.1
NB	52.8	45.8	100.0	40.1

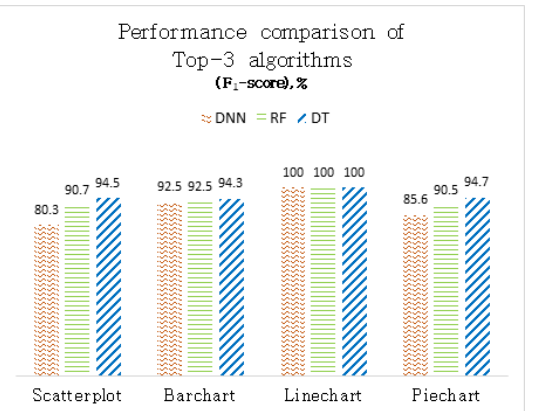

그림 10. 상위 3개 알고리즘의 성능 비교
Fig. 10. Performance comparison of the Top-3 algorithms

표 5. 시각화 종류별 Decision Tree모델의 분류 성능
Table 5. Classification performance of decision model by visualization type

	Decision Tree (%)		
	Precision	**Recall**	**F_1-score**
S	94.3	94.8	94.5
B	94.2	95.2	94.3
L	100.0	100.0	100.0
P	94.3	95.1	94.7

밀도(precision), 재현율(recall), $F_1-score$ 값을 담고 있으며, 시각화 종류별 분류(추천) 성능을 확인할 수 있다. 성능 평가 지표인 $F_1-score$의 값은 각각의 시각화 종류별로 보았을 때, 산점도(S)는 94.5%, 막대그래프(B)는 94.3%, 라인그래프(L)는 100.0%, 파이그래프(P)는 94.7%이다.

DeepEye 연구에서도 Decision Tree 알고리즘이 가장 좋은 성능을 보였으며, DeepEye에서 발표한 Decision Tree의 분류 성능과 본 연구에서 수행한 실험 결과와 비교하였을 때, 산점도는 1.5%, 막대그래프는 1.2%, 라인그래프는 0.5% 정도의 분류 성능이 향상됨을 확인하였다(그림 11. 참조). 이는 우리의 메타 특징 공학을 통해 기존의 DeepEye의 성능보다 개선되었으며, 데이터 시각화의 자동화가 충분히 가능할 수 있음을 보여준다.

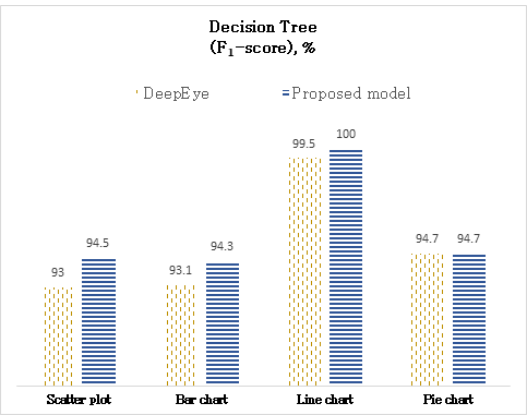

그림 11. 제안 기법과 비교 기법의 성능 비교
Fig. 11. Performance comparison with propsed model

V. 결론 및 향후연구

본 논문은 자동 데이터 시각화를 위한 메타성 특징 데이터(meta data)를 설계하고, 결과적으로 정확도가 높은 시각화 추천 모델을 구축하기 위한 메타특징 공학 과정을 소개하였다. 기존 DeepEye에서 제안한 특징공학은 최소한의 특징변수 14개를 추출하여 실험을 진행했지만, 우리는 유의미한 시각화 패턴을 효과적으로 학습하기 위한 특징변수 27개로 확장하여 기존 모델보다 분류(추천) 성능을 높일 수 있었다. 자동 데이터 시각화 결과의 가치성이 높기 위해서는 그것이 유의미한 분포, 패턴 등의 정보를 시각적으로 표출해야 한다. 본 연구를 통해 우리는 주어진 데이터에 내재된 양질의 메타 특징 정보가 합당한 수준의 시각화 결과를 자동 추천할 수 있는 토대가 됨을 확인하였다.

향후 연구로서 시각화의 유형을 늘리고, 데이터 프로파일링 기반 메타특징의 추가 확장을 통해 자동시각화 시스템의 실현성을 높일 것이다.

References

[1] J. Y. Byun and Y. B. Park, "A guiding system of visualization for quantitative bigdata based on user intention," *KIPS Trans. Softw. and Data Eng.*, vol. 5, no. 6, pp. 261-266, Jun. 2016.

[2] A. Kirk, *Data visualization: A successful design process*, Packt Publishing Ltd., 2012.

[3] B. Fry, *Visualizing Data*, O'REILLY Publishing Ltd., pp. 14-15, 2007.

[4] Y. Luo, X. Qin, N. Tang, and G. Li, "DeepEye towards automatic data visualization," *2018 IEEE 34th ICDE*, Paris, France, Apr. 2018.

[5] M. O. Ward, G. Grinstein, and D. Keim, *Interactive data visualization: Foundations, techniques, and applications*, AK Peters/ CRC Press, pp. 81-86, 2015.

[6] M. Khan and S. S. Khan, "Data and information visualization methods, and interactive mechanisms: A survey," *IJCA*, vol. 34, no. 1, Nov. 2011.

[7] Adobe Flex 3, "*Advanced Data Visualization Developer*(2008)," pp. 40, 66-69, Sep. 2018, http://few.vu.nl/~eliens/assets/flex3/datavis_flex3.pdf

[8] S. Few and P. Edge, “Save the pies for dessert,” *Visual Business Intell. Newsletter*, 2007.
[9] S. Wright, “Correlation and causation,” *J. Agricultural Res.*, vol. XX, no. 7, pp. 557-563, Jan. 1921.
[10] D. Cournapeau, *Scikit-learn*(2007), Mar. 01. 2018, https://scikit-learn.org/stable/
[11] F. Chollet, *Keras*(2015), Sep. 01, 2018, https://keras.io/

최 희 원 (Hee-won Choi)

2015년 2월 : 덕성여자대학교 정보통계학과 이학사
2018년~현재 : 서울시립대학교 전자전기컴퓨터공학부 석사과정
<관심분야> 머신러닝, 자동시각화, 추천시스템, 텍스트마이닝
[ORCID:0000-0001-7542-5853]

김 한 준 (Han-joon Kim)

1994년 2월 : 서울대학교 계산통계학과 이학사
1996년 2월 : 서울대학교 전산과학과 이학석사
2002년 2월 : 서울대학교 컴퓨터공학부 공학박사
2002년~현재 : 서울시립대학교 전자전기컴퓨터공학부 정교수
<관심분야> 텍스트마이닝, 머신러닝, 온톨로지, 정보검색, 데이터베이스, 빅데이터 기술
[ORCID:0000-0003-4510-5685]