ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 September: 22(3): 143~163 http://dx.doi.org/10.13088/jiis.2016.22.3.143

인구통계특성 기반 디지털 마케팅을 위한 클릭스트림 빅데이터 마이닝

박지애
국민대학교 데이터사이언스학과
(*lucky0240@naver.com*)

조윤호
국민대학교 경영대학 경영학부
(*www4u@kookmin.ac.kr*)

인구통계학적 정보는 디지털 마케팅의 핵심이라 할 수 있는 인터넷 사용자에 대한 타겟 마케팅 및 개인화된 광고를 위해 고려되는 가장 기초적이고 중요한 정보이다. 하지만 인터넷 사용자의 온라인 활동은 익명으로 행해지는 경우가 많기 때문에 인구통계특성 정보를 수집하는 것은 쉬운 일이 아니다. 정기적인 설문 조사를 통해 사용자들의 인구통계특성 정보를 수집할 수도 있지만 많은 비용이 들며 허위 기재 등과 같은 위험성이 존재한다. 특히, 모바일 환경에서는 대부분의 사용자들이 익명으로 활동하기 때문에 인구통계특성 정보를 수집하는 것은 더욱 더 어려워지고 있다. 반면, 인터넷 사용자의 온라인 활동을 기록한 클릭스트림 데이터는 해당 사용자의 인구통계학적 정보에 활용될 수 있다. 특히, 인터넷 사용자의 온라인 행위 특성 중 하나인 페이지뷰는 인구통계학적 정보 예측에 있어서 중요한 요인이 된다. 본 연구에서는 기존 선행 연구를 토대로 클릭스트림 데이터 분석을 통해 인터넷 사용자의 온라인 행위 특성을 추출하고 이를 해당 사용자의 인구통계학적 정보 예측에 사용한다. 또한, 1)의사결정나무를 이용한 변수 축소, 2)주성분분석을 활용한 차원축소, 3)군집분석을 활용한 변수 축소의 방법을 제안하고 실험에 적용함으로써 많은 설명변수를 이용하여 예측 모델 생성 시 발생하는 차원의 저주와 과적합 문제를 해결하고 예측 모델의 정확도를 높이고자 하였다. 실험 결과, 범주의 수가 많은 다분형 종속변수에 대한 예측 모델은 모든 설명변수를 사용하여 예측 모델을 생성했을 때보다 본 연구에서 제안한 방법론들을 적용했을 때 예측 모델에 대한 정확도가 향상됨을 알 수 있었다. 본 연구는 클릭스트림 분석을 통해 추출된 인터넷 사용자의 온라인 행위는 해당 사용자의 인구통계학적 정보 예측에 활용 가능하며, 예측된 익명의 인터넷 사용자들에 대한 인구통계학적 정보를 디지털 마케팅에 활용 할 수 있다는데 의의가 있다. 또한, 제안 방법론들을 통해 어느 종속변수에 대해 어떤 방법론들이 예측 모델의 정확도를 개선하는지 확인하였다. 이는 추후 클릭스트림 분석을 활용하여 인구통계학적 정보를 예측할 때, 본 연구에서 제안한 방법론을 사용하여 보다 높은 정확도를 가지는 예측 모델을 생성 할 수 있다는데 의의가 있다.

주제어 : 빅데이터, 클릭스트림 분석, 인구통계학 정보, 온라인 행위, 분류분석, 변수차원 축소

논문접수일 : 2016년 8월 17일 논문수정일 : 2016년 9월 19일 게재확정일 : 2016년 9월 24일
원고유형 : 일반논문 교신저자 : 조윤호

1. 서론

인터넷 발달과 더불어 PC 및 모바일을 사용한 온라인 활동이 활발해짐에 따라 타겟 마케팅, 개인화된 광고 등의 수요가 증가하고 있다. 성별, 연령 등을 포함하는 인구통계특성 정보는 타겟 마케팅 및 개인화된 광고를 위해서 고려되어야 할 가장 기초적이고 중요한 정보이다. 하지만 인터넷 사용자의 온라인 활동은 익명으로 행해지는 경우가 많기 때문에 인구통계특성 정보를 수

집하기는 어렵다. 회원가입과 같은 사용자들의 자발적 등록을 통해서 해결될 수 있지만 이는 많은 사람들이 사용하는 특정 웹사이트에서만 가능한 접근 방법이다. 또한 정기적인 설문 조사를 통해 사용자들의 인구통계특성 정보를 수집할 수도 있지만 많은 비용이 들며 허위 기재 등과 같은 위험성이 존재한다(De Bock and Van den Poel, 2009). 특히 모바일 환경에서는 대부분의 사용자들이 익명으로 활동하기 때문에 인구통계특성 정보를 수집하는 것은 더욱 더 어려워지고 있다.

반면 인터넷 사용자들의 온라인 활동 내역을 기록한 클릭스트림 데이터(clickstream data)는 해당 사용자의 인구통계특성 예측에 활용될 수 있다. Eleonora(2013)의 연구에 따르면, 클릭스트림 데이터로부터 추출된 인터넷 사용자의 온라인 행위 정보와 인구통계특성 정보는 상관성이 있으며 구체적으로 사용자들의 웹사이트 관심사 및 선호 시간대에 따른 페이지뷰(pageviews)는 성별과 연령을 예측하는데 중요한 요인이 된다. 이처럼 국외에는 클릭스트림 분석을 토대로 온라인 행위 정보를 추출하고 이를 이용한 인구통계특성 예측에 관한 연구가 몇 차례 진행되었지만(De Bock and Van den Poel, 2009; Eleonora, 2013; Kim, 2011; Murray et al., 2000) 국내에는 거의 없는 실정이다. 실제로 인터넷 사용자의 온라인 행위 정보에 관한 국내 연구 동향을 살펴보면 주로 인터넷 사용자의 인구통계특성에 따른 온라인 행위를 확인하거나, 온라인 행위 정보를 이용하여 온라인 상점에서 인터넷 사용자의 구매 의도를 파악하거나, 모바일 디바이스에서 발생하는 특징적 단어와 같은 텍스트 정보만을 활용하여 인터넷 사용자의 인구통계특성을 확인하는 등 클릭스트림 데이터로부터 추출된 인터넷 사용자의 온라인 행위 정보를 이용하여 해당 사용자의 인구통계특성을 예측하는 시도는 없었다(Han et al., 2012; Kim et al., 2016).

본 연구의 목적은 다음과 같다. 첫째, 클릭스트림 데이터에서 분석을 토대로 인터넷 사용자들의 온라인 행위 정보를 추출하고 이를 이용하여 해당 사용자의 인구통계특성을 예측하고자 한다. 둘째, 예측 모델 생성 시 발생 가능한 차원의 저주(curse of dimensionality) 및 과적합(overfitting) 문제를 해결하기 위해 여러 변수축소 방법들을 적용하고 각각의 예측 모델의 성능을 비교하고자 한다.

2. 이론적 배경 및 관련 연구

2.1 클릭스트림 데이터

디지털 기술의 급속한 성장으로 인터넷 사용자의 온라인 활동에 대한 중요성은 대두고 있다. 그럼에도 불구하고 인터넷 사용자들의 온라인 행위에 관한 몇 가지 의문점은 존재한다. 인터넷 사용자가 그들의 인구통계특성에 따라 얼마나 자주 건강 관련 정보에 접근하는지, 웹 트래픽의 대부분을 차지하는 사용자들은 일반적인 사용자와 온라인 행위에서 어떤 차이가 존재하는지 등을 예로 들 수 있다. 이러한 의문점에 대한 해결은 정보격차에 대한 대응에서부터 마케팅과 광고 전략 개선까지 다양한 시사점을 제공할 수 있다(Goel et al, 2012).

일반적으로 웹서핑(web surfing)으로 표현되는 온라인 행위는 인터넷 사용자가 정보탐색, 온라인 구매 등의 활동을 위해 웹사이트를 방문하는 행위를 의미한다. 이러한 인터넷 사용자의 온라

인 행위 정보는 사용자가 하나 또는 그 이상의 웹사이트를 방문한 경로를 기록한 클릭스트림 데이터를 이용하여 보다 객관적이고 정확하게 확인할 수 있으며 클릭스트림 데이터는 설문조사에서 얻을 수 없는 정보까지도 포함하고 있다는 장점을 가지고 있다. 다시 말해, 클릭스트림 데이터에는 인터넷 사용자가 어떤 사이트를 방문했는지, 특정 웹사이트를 얼마나 자주 방문했는지 등에 대한 일련의 온라인 활동들이 기록되어있다(Lourenco et al., 2011).

또한, 클릭스트림 데이터는 웹사이트를 운용하는 업체가 가진 서버의 로그에서 웹로그(web log) 데이터 형태로 얻어지거나 사용자 컴퓨터에 소프트웨어를 설치하여 패널(panel)들의 방문 정보를 수집하는 기관으로부터 패널 데이터(panel data)의 형태로 얻어진다. 웹로그 데이터 형태의 클릭스트림 데이터에는 방문한 웹페이지 주소, 접속날짜와 시간, 세션 수 등과 같은 인터넷 사용자의 온라인 활동 기록에 대한 정보들이 포함되며 패널 데이터 형태의 클릭스트림 데이터에는 온라인 활동 기록뿐만 아니라 사용자의 인구통계특성 정보까지 포함된다. 클릭스트림 데이터는 인터넷 사용자들에 대해 다양한 웹사이트 간의 이동경로를 추적하여 이동경로 관계를 파악하거나 이동경로 모형화에 활용되며(Awad et al., 2006; Bucklin. R, 2002; Park and Fader, 2004), 온라인 쇼핑몰 사용자의 다음 방문 시 구매여부를 예측하는(Moe and Wendy, 2003) 등 다양한 분야의 연구에 활용되어 왔다.

2.2 온라인 행위를 이용한 인구통계특성 예측

웹사이트를 운용하는 업체들은 자사의 웹사이트를 이용하는 사용자들에 대한 인구통계학적 분포와 사용자 개개인에 대한 정보 획득을 위해 노력한다. 그럼에도 불구하고 대부분의 업체들은 익명의 사용자들의 웹사이트 방문 횟수를 토대로 간단한 통계학적 분석만 할 수 있다는 한계점을 가진다. 하지만 클릭스트림 분석을 통해 추출한 온라인 행위 정보를 이용하여 인터넷 사용자들에 대한 인구통계특성은 추론 가능하며 이에 관한 연구 동향은 <Table 1>과 같다.

인터넷 사용자의 온라인 행위 정보를 이용하여 인구통계특성을 예측한 연구의 시초는 Murray et al.,(2000)의 연구로 인터넷 사용자들이 검색엔진과 방문 웹페이지 등에서 검색한 검색 키워드를 이용하여 해당 사용자들의 성별과 연령 등을 예측했다. 이때 사용한 알고리즘은 LSA와 신경망이며 모델 성능 측정을 위해 향상도(lift)를 사용하였다. Baglioni et al.,(2003)의 연구는 인터넷 사용자가 방문한 URL에 포함된 단어와 방문 URL의 구조로부터 구문과 의미에 대한 정보를 추출하여 해당 사용자의 성별을 예측하는데 사용하였다. 예측 모델을 학습하기 위한 분류 기법으로 의사결정나무(decision tree)의 알고리즘 중 하나인 C4.5와 k-최근접 이웃(k-nearest neighbors)을 사용하였다. 또한, 나이브 베이즈(naive bayes)를 기준으로 향상도(lift)를 이용하여 각각의 예측 모델의 성능을 측정하였다. 실험 결과, 향상도는 10.2%로 다소 낮게 측정되었으며 연구자는 이를 예측 모델 학습 시 지나치게 일반적인 알고리즘을 사용했기 때문이라고 설명했다. 성별과 연령 예측에 대한 또 다른 연구로 Jones et al.(2007)에서는 인터넷 사용자들의 검색 키워드만을 설명변수로 사용하고 SVM(support vector machine) 분류 기법을 적용하여 해당 사용자의 성별을 예측하였다.

De Bock and Van den Poel(2009)는 클릭스트

〈Table 1〉 Literature review Existing methods for predicting demographics

	Independent Variable	Dependent Variable	Algorithms	Performance	Author and year
A	· Search Keyword · Text Information for webpages visited	· Gender(2) · Age Under 18(2) · Age 18~34(2) · Age 35~54(2) · Age 55+(2) · Income Over $50,000(2) · Marital status(2) · Some College Education(2) · Presence of Children(2)	· LSA · Neural Network	Lift	Murray and Durrell, 2000
B	· Information of words, syntax and semantics for URLs visited	· Gender(2)	· C4.5 · k-Nearest Neighbors		Baglioni et al., 2003
C	· Search Keyword · Text Information for webpagess visited	· Gender(2)	· SVM	Accuracy	Jones et al., 2007
D	· Frequency and Intensity for time, day and month · Variety of websites visited	· Gender(2) · Age(6) · Education(5) · Job(10)	· Random Forest		De Bock and Van den Poel, 2009
E	· Preferences of website · Ad exposures · Ad clicks · Variety of websites visited · Day and time of visit frequency	· Gender(2) · Age Under 35(2) · Some College Education(2) · Income Over £50,000(2) · Marital status(2) · Presence of Children(2)	· Logistic Regression		Eleonora, 2013

* The figures in parentheses are the number of categories for each variable.

림 데이터에서 추출한 온라인 행위 정보를 웹사이트에 대한 방문빈도(visit frequency)와 체류시간 및 페이지뷰로 설명되는 방문강도(visit intensity)로 구분하여 설명변수로 사용하였다. 의사결정나무 알고리즘 중 하나인 랜덤포레스트(random forest)를 이용하여 사용자들의 성별, 연령, 최종학력, 직업을 예측하였고 예측 모델의 성능 평가를 위해 mAUC와 정확도(accuracy)를 사용하였다. Eleonora(2013)는 클릭스트림 데이터를 사용하여 광고 노출 수, 광고 링크를 클릭한 수, 방문 웹사이트 다양성, 시간대 및 요일별 방문횟수에 대한 온라인 행위 정보를 추출하고 이를 설명 변수로 사용하였다. 또한, 로지스틱 회귀(logistic regression)를 이용하여 이분형(binary)으로 측정된 사용자의 인구통계특성을 예측하고 모델의 평가 척도로 정확도를 사용하였다.

De Bock and Van den Poel(2009)의 연구를 제외한 대부분의 연구에서 인터넷 사용자의 인구통계특성을 이분형 종속변수로 설정하여 예측 모델을 생성하였다. 일반적으로 다분형보다 이분형 종속변수에 대한 예측 모델의 정확도가 우

수하지만 온라인과 오프라인 마케팅 모두 각각의 고객에 대한 타겟 마케팅 및 개인화 서비스의 수요가 증가하고 있으며 고객의 다양한 정보를 활용해 가장 확실한 가치를 찾는 핀셋 마케팅 또는 극세분화 마케팅이 떠오르는 현재 시점에서 인터넷 사용자의 인구통계특성을 이분형 종속변수로 설정하여 예측하는 것은 실무적 활용성이 현저히 낮을 것이라 판단된다. 따라서 본 연구는 De Bock and Van den Poel(2009)의 연구를 중점으로 앞의 모든 선행연구를 활용하여 클릭스트림 데이터에서 인터넷 사용자의 온라인 행위 정보를 추출하고 이를 사용하여 이분형 및 다분형으로 표현되는 해당 사용자의 인구통계특성을 예측하고자 한다.

2.3 관련 분석 기법

2.3.1 SVM(support vector machine)

SVM(support vector machine)은 선형, 비선형을 가리지 않고 데이터를 분류할 수 있는 기법이다. 기본 원리는 목표값을 분류할 때 기준이 되는 최적의 분리경계면(hyperplane)을 찾는 것을 목적으로 한다. SVM은 일반적으로 이진 분류(binary classification) 알고리즘으로 사용되나 다분형 SVM(multi-class support vector machine)을 이용하여 다항 분류(multinominal classification)도 가능하다. Foody and Marther(2004)에 의하면 여러 개의 이분형 SVM을 결합하여 다분형 SVM을 구성하는 방식으로 one-against-all과 one-against-one이 있다. SVM은 훈련시간이 다른 분류 기법에 비해 느리지만, 정확성이 매우 뛰어나고 복잡한 비선형 모델까지도 구축 가능하며 과적합 문제의 발생 가능성 또한 낮다(Kim and Ahn, 2015).

2.3.2 신경망(neural network)

신경망(neural networks)은 생물학적 신경망에서 영감을 얻은 알고리즘으로 모델의 성능이 높은 분류 기법으로 알려져 있다. 가장 일반적인 신경망 모형은 다층 퍼셉트론(multi-layer perceptron, MLP) 구조로 각 뉴런이 서로 연결되어있는 것이 특징이다(Rumelhart, 1986). 신경망 모델은 입력값이 많을수록 즉, 조정해야 할 가중치가 많을수록 모델의 복잡도(complexity)는 증가하고 모델 학습 시 차원의 저주 및 과적합 문제가 발생할 가능성이 높아진다. 신경망 모델의 차원의 저주 및 과적합 문제를 해결하기 위해서는 다량의 데이터를 이용하거나 적절한 입력변수를 선택하여 신경망 모델을 생성해야 한다.

2.3.3 로지스틱 회귀(logistic regression)

로지스틱 회귀(logistic regression)는 일반적으로 이항 로지스틱 회귀(binomial logistic regression)를 의미하며 이분형 종속변수를 분류하기 위해 사용되는 통계적 기법이지만 다항 로지스틱 회귀(multinomial logistic regression)를 이용하여 범주가 3개 이상인 다분형 종속변수의 분류도 가능하다.

2.3.4 의사결정나무(decision tree)

의사결정나무(decision tree)는 해석이 용이하고 구현에 어려움이 없어 흔히 사용된다. 의사결정나무가 활용될 수 있는 응용분야는 크게 4가지로 구분가능하며 세분화(segmentation), 분류, 예측, 차원축소 및 변수선택(dimensional reduction and variable selection)이 이에 해당한다. 특히 차원축소 및 변수선택 방법은 설명변수의 수가 매우 많을 때 예측 모델의 성능 향상을 위해 의사

결정나무에서 예측 변수 중요도를 추출하여 상대적으로 종속변수 예측에 큰 영향을 미치는 설명변수들을 선택하는 방법이다(Lee and Lee, 2003; Genuer et al., 2010).

2.3.5 주성분분석(principal component analysis)

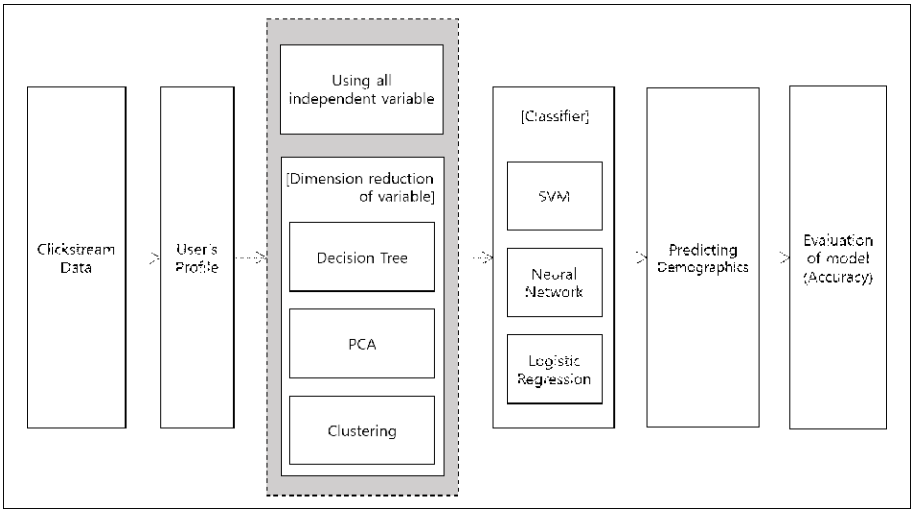
주성분분석(principal component analysis)은 고차원 데이터 집합을 저차원으로 축소시키는 대표적인 차원축소 방법으로 상관관계가 높은 변수들의 선형결합을 통해 p개의 변수들을 $m(m < p)$ 개의 주성분으로 변환시킨다. 본 연구에서는 예측 모델 학습 시 차원의 저주 및 과적합 문제를 해결하고자 주성분분석을 사용하였다.

2.3.6 군집분석(cluster analysis)

군집분석(cluster analysis)은 각 객체의 유사성을 측정하고 유사성이 높은 객체끼리 집단을 형성하는 방법이며 주로 세분화, 이상치 탐지(anomaly detection) 등에 사용된다(Cho and Park, 2012). 본 연구에서는 변수축소의 방법으로 군집분석을 활용하여 여러 개의 연속형 변수를 하나의 명목형 변수로 축소시켰다.

3. 제안 방법론

본 연구의 인터넷 사용자의 인구통계특성 예측에 관한 실험은 <Figure 1>과 같은 프로세스를 가진다. 먼저 클릭스트림 데이터에서 추출되어 인터넷 사용자의 인구통계특성을 예측할 때 사용되는 데이터 집합을 사용자 프로파일로 정의한다. 사용자 프로파일은 선행 연구를 토대로 온라인 행위 정보에 관한 64개의 변수와 인구통계특성 정보에 관한 5개의 변수로 구성된다. 사용자 프로파일 중 온라인 행위 정보에 관한 64개의 모든 변수를 이용하여 예측 모델을 학습할 경우 차원의 저주 및 과적합 문제가 발생할 가능성이 높다. 따라서 본 연구에서는 차원의 저주 및 과



〈Figure 1〉 Process of predictive model building

적합 문제를 해결하기 위해 (1)의사결정나무를 이용한 변수축소, (2)주성분분석을 이용한 변수축소, (3)군집분석을 활용한 변수축소 방법을 제안한다.

3.1 사용자 프로파일(profile) 생성

본 연구에서는 선행연구를 토대로 클릭스트림 데이터로부터 사용자의 온라인 행위 정보와 인구통계특성 정보를 추출하여 사용자 프로파일을 생성하였다. 본 연구의 실험에서 사용되는 사용자 프로파일은 <Table 2>에 요약되어있다. 사용자 프로파일의 인터넷 사용자별 선호사이트는 Eleonora(2013)의 연구에서 사용된 웹사이트 방문 선호도의 변수와 같은 맥락으로 사용자가 어떤 카테고리의 웹사이트에 자주 접속하는지를 나타낸다. 웹사이트를 22개의 카테고리로 구분하고 각각의 카테고리별 사용자의 페이지뷰 비율과 카테고리에 대한 변동계수(coefficient of variance)로 표현했고 22개의 카테고리는 건강/의학, 게임, 금융/부동산, 뉴스미디어 등으로 분류하였다. 카테고리별 페이지뷰 비율은 0과 1사이의 값을 가지며 연속형 변수로 표현되며 큰 값을 가질수록 해당 웹사이트 카테고리에 대한 방문 선호도가 높다는 것을 의미한다. 또한, 사용자별로 변수를 생성하였기 때문에 사용자 1명당 카테고리별 페이지뷰 비율의 합은 1이다. 카테고리에 대한 변동계수는 클릭스트림 데이터를 수집한 1년의 기간 동안 사용자가 웹사이트 카테고리 22개에 대해 얼마나 다양하게 이용을 하였는지에 대한 것으로 해당 값들에 대한 표준편차를 평균으로 나눈 값을 의미하며 값이 클수록 특정 웹사이트 카테고리에 집중 방문하며 해당 웹사이트 카테고리를 선호하는 것을 의미한다. 다음으로 웹사이트 사용 패턴은 여러 선행연구(De Bock and Van den Poel, 2009; Eleonora, 2013; Goel et al., 2012)를 토대로 사용자가 주로 어떤 시간대, 어떤 요일, 어떤 월에 접속하는지에 대한 것이다. 총 페이지뷰 횟수, 총 방문 일수, 시간대별 페이지뷰 비율, 요일별 페이지뷰 비율, 월별 페이지뷰 비율, 시간대에 대한 변동계수, 요일에 대한 변동계수, 월에 대한 변동계수가 웹사이트 사용 패턴에 해당된다. 총 페이지뷰 횟수는 사용자에 대한 1년 동안 페이지뷰 횟수의 총합을 나타내며, 총 방문 일수는 1년 365일 중 웹사이트에 접속한 총 일 수이다. 시간대별 페이지뷰 비율, 요일별 페이지뷰 비율, 월별 페이지뷰 비율은 0과 1사이의 값을 가지며 연속형 변수로 표현되며 큰 값을 가질수록 해당 시간대, 요일, 월에 대한 방문 선호도가 높다는 것을 의미한다. 여기서 시간대은 24시간을 00~05시, 06~11시, 12~17시, 18~23시 4개로 구분하여 사용하였고 요일과 월은 7개의 요일과 12개의 월을 사용하였다. 시간대, 요일, 월에 대한 변동계수는 클릭스트림 데이터를 수집한 1년의 기간 동안 사용자가 각각의 시간대, 요일, 월에 대해 얼마나 다양하게 이용을 하였는지에 대한 것으로 해당 값들에 대한 표준편차를 평균으로 나눈 값을 의미하며 값이 클수록 특정 시간대, 요일, 월에 대한 온라인 활동이 활발한 것을 의미한다.

검색행위는 Murray and Durrell(2000)와 Jones et al.(2007)의 연구에서 사용자의 인구통계특성을 예측할 때, 중요한 변수가 됨을 확인하였다. 또한, Gallagher and Parsons(1997)는 인터넷 사용자의 인구통계특성은 정보탐색에서 차이를 보이며 특히 사용자의 연령이 높을수록 더 많은 검색 키워드를 사용하는 것을 확인하였다. 따라서 본 연구에서 사용된 검색행위는 인터넷 사용자의

〈Table 2〉 Summary of User's Profile

		Variable description and the number of variables		Author and year
Behavior	Preferences of websites	Ratio of pagevews for website category	22	De Bock and Van den, 2009 Goel et al, 2012
		Coefficient of variation for website category	1	
	Usage Pattern	Total number of pageviews for website category	1	De Bock and Van den, 2009 Eleonora, 2013 Goel et al., 2012
		The total number of days to visit	1	
		Ratio of pagevews for time	4	
		Ratio of pagevews for day	7	
		Ratio of pagevews for month	12	
		Coefficient of variation for time	1	
		Coefficient of variation for day	1	
		Coefficient of variation for month	1	
	Search Behavior	Total number of search keywords	1	Gallagher and Parsons, 1997 Jones et al., 2007 Murray and Durrell, 2000
	Interest	Ratio of pagevews for news website category	12	Yoonjin Hyun et al. ,2015
Demographics		Gender(2)	1	Baglioni et al., 2003 De Bock and Van den, 2009 Eleonora, 2013 Murray and Durrell, 2000 Jones et al., 2007
		Age(5)	1	De Bock and Van den, 2009 Eleonora, 2013 Murray and Durrell, 2000
		Marital Status(2)	1	Eleonora, 2013 Murray and Durrell, 2000
		Residence(13)	1	Eleonora, 2013
		Job(20)	1	De Bock and Van den, 2009 Eleonora, 2013

* The figures in parentheses are the number of categories for each variable.

총 검색키워드 수로 표현하였고 총 검색키워드 수는 사용자가 1년 동안 포털 사이트에서 검색한 검색키워드의 총 개수를 의미한다.

마지막으로 사용자의 관심사를 측정하기 위한 변수로 뉴스사이트 카테고리별 페이지뷰 비율을 사용하였다. 뉴스기사의 독자들은 그들의 관심사에 따라 특정 뉴스기사를 선택하고(Poindexter et al., 2001) 사용자의 인구통계특성에 따라 선호

하는 뉴스기사에 대한 주제가 다르며 이용량에도 차이가 있다(Ban and Kwon, 2007). 또한 Choi et al. (2015)의 연구에서도 인터넷 사용자의 관심사를 측정하기 위해 뉴스기사 접속 기록을 사용하였다. 따라서 본 연구에서는 사용자의 관심사를 측정하기 위한 척도로 뉴스사이트의 카테고리별 페이지뷰 비율을 사용하였고 뉴스사이트는 속보, 정치, 경제 등 총 12개의 카테고리로 분류하였다. 뉴스사이트의 카테고리별 페이지뷰 비율은 0과 1사이의 값을 가지며 연속형 변수로 표현되며 큰 값을 가질수록 해당 카테고리의 뉴스사이트에 대한 선호도가 높다는 것을 의미한다. 마지막으로 사용자의 인구통계특성에 관한 정보는 성별, 나이, 혼인여부, 거주지, 직업 정보를 포함한다. 성별은 남자와 여자로 나뉘며, 나이는 10대 미만, 20대, 30대, 40대, 50대 이상 총 5개의 그룹으로 분류되며 혼인여부는 기혼과 미혼으로 나뉜다. 또한 거주지는 서울, 경기, 충청 등 총 13개 클래스로 분류되며 직업은 방송/예술/스포츠, 자영업, 학생 등 20개 클래스로 나뉜다. 이에 따라 본 연구에 사용된 사용자 프로파일은 온라인 행위 정보에 관한 64개의 변수와 인구통계특성 정보에 관한 5개의 변수로 구성된다.

3.2 변수축소 방법

3.2.1 의사결정나무를 이용한 변수축소

예측 변수 중요도는 매우 많은 설명변수 중에서 상대적으로 종속변수에 큰 영향을 미치는 변수를 구분하는 방법이다. 설명변수의 수를 줄임으로써 차원의 저주 및 과적합 문제를 피하고 예측 모델의 정확도를 높일 수 있으며 예측 변수 중요도는 의사결정나무 분석을 통해 확인 가능하다(Lee and Lee, 2003; Genuer et al., 2010).

3.2.2 주성분분석을 이용한 변수축소

주성분분석은 고차원 데이터의 차원을 축소하기 위한 방법이다. 주성분분석으로부터 추출된 주성분은 기존 변수들의 선형결합으로 구성되며 소수의 주성분으로써 데이터를 설명할 수 있기 때문에 고차원 데이터의 차원축소에 효과적으로 활용되고 있다(Boutsidis et al., 2008). 따라서 본 연구에서는 변수축소 방법 중 하나로 주성분분석을 이용하였다.

3.2.3 군집분석을 활용한 변수축소

본 연구에서는 군집분석을 활용한 데이터의

Category 22

user_id	Finance/ Real Estate	News media	Entertainment	...	Travel	Internet/ Computer	Community
1	0.00	0.21	0.11		0.00	0.41	0.22
2	0.01	0.03	0.01		0.00	0.37	0.10
3	0.00	0.03	0.01		0.01	0.69	0.03
4	0.02	0.18	0.00		0.00	0.56	0.17
5	0.05	0.25	0.00		0.00	0.55	0.05
⋮						⋮	
4997	0.09	0.14	0.00		0.00	0.25	0.19
4998	0.08	0.03	0.02		0.02	0.56	0.46
4999	0.00	0.06	0.00		0.02	0.36	0.10
5000	0.02	0.05	0.08		0.00	0.83	0.11

Cluster analysis →

user_id	CT_Cluster
1	cluster1
2	cluster1
3	cluster3
4	cluster5
5	cluster3
⋮	
4997	cluster1
4998	cluster2
4999	cluster4
5000	cluster4

〈Figure 2〉 An example of dimension reduction using clustering analysis

변수축소 방법을 새롭게 제시한다. 군집분석을 활용하여 여러 개의 연속형 변수를 하나의 범주형 변수로 축약하는 방법이다. 이해를 돕기 위해 군집분석을 활용한 변수축소 방법의 예시를 <Figure 2>에 제시하였다. <Figure 2>에는 연속형으로 표현된 웹사이트 카테고리별 페이지뷰 비율 변수 22개를 군집분석을 통해 하나의 범주형 변수로 축소시키는 것에 대한 예시이다.

4. 실험 및 평가

4.1 실험 데이터

본 연구의 실험에 사용된 데이터는 국내의 한 인터넷 사이트 순위 분석 전문업체로부터 패널 5,000명에 대해 2012년 7월 1일부터 2013년 6월 30일까지 1년 동안의 온라인 활동 기록을 제공받은 패널 데이터 형태의 클릭스트림 데이터이다. 해당 데이터는 패널 5,000명의 인구통계특성에 관한 정보 5항목(성별, 연령, 혼인여부, 거주지, 직업)과 해당 패널에 대한 해당 기간 동안의 온라인 활동 기록 16,962,705건에 대한 상세항목 6항목(방문 사이트, 접속 날짜, 접속 시각, 페이지뷰, 검색키워드)으로 구성된다. 클릭스트림 데이터를 활용하여 생성한 사용자 프로파일은 인터넷 사용자(패널)를 기준으로 5,000개의 행(rows)과 온라인 행위 정보에 대한 64개의 열(columns)과 인구통계특성 정보에 대한 5개의 열로 구성되어 있다. 예측모델 생성을 위해 온라인 행위 정보에 관한 64개의 열을 설명변수로 사용하고 인구통계특성 정보에 대한 5개의 열을 종속변수로 사용한다.

4.2 실험 방법

실험을 위해 사용자 프로파일은 예측 모델의 학습을 위한 훈련용 데이터와 예측 모델의 성능 측정을 위한 검증용 데이터로 분할하여 사용하였다. 또한, 본 연구결과에 대한 신뢰성을 높이기 위해 각각의 분류 기법에 대해 5-fold cross-validation을 사용하였다. 사용자 프로파일을 사용자를 기준으로 5개의 동일한 크기의 그룹(fold)으로 분할하여 4개의 그룹은 학습용 데이터로, 나머지 1개의 그룹은 검증용 데이터로 사용한다. 이 과정을 중복 없이 5번 반복하여 모델의 평균 성능을 사용한다. 다시 말해 5,000개의 행으로 이루어진 사용자 프로파일을 훈련용 데이터와 검증용 데이터로 분할한다. 이때, 훈련용 데이터의 크기가 64개의 설명변수를 이용하여 예측 모델을 학습하기에 매우 적은 양으로 차원의 저주 및 과적합 문제가 발생할 수 있다. 따라서 본 연구는 차원의 저주 및 과적합 문제를 해결하기 위해서 총 3가지의 변수축소 방법을 적용한 예측 모델을 모두 생성하고 정확도(accuracy)를 이용하여 예측 모델의 성능을 비교 평가한다. 정확도는 모델의 성능을 단 하나의 수치로 표현되며 측정하기가 쉽기 때문에 분류 모델의 평가 척도로 널리 사용되는 모델 성능 측정방법이다. 또한 본 연구에서는 실험을 위해 데이터마이닝 도구인 SPSS Modeler 17.0을 사용하였다.

4.3 실험 결과

본 연구의 실험에 사용된 분류 기법은 SVM, 신경망, 로지스틱 회귀이다. SVM을 이용한 예측 모델 학습 시 커널 함수로서 RBF(radial basis function)함수를 사용하였다. 신경망을 이용한 예

측 모델은 다중 레이어 퍼셉트론 구조를 기반으로 학습되었으며 훈련용 데이터에 대해 신경망 모델을 기저 학습법으로 사용하는 배깅(bagging)을 적용하였다. 다음은 본 연구에서 제안한 각각의 변수축소 방법에 대한 실험 결과이다.

4.3.1 전체 설명변수 사용

사용자 프로파일에서 64개의 온라인 행위 변수들을 모두 사용하여 각각의 인구통계특성을 예측한 결과는 다음 <Table 3>와 같다. <Table 3>에 기재된 예측 모델의 정확도는 5-fold cross-validation을 적용하여 평균 정확도를 계산한 것으로 괄호 안에 있는 수치는 표준편차 값이다.

〈Table 3〉 Accuracy for using all variables

	SVM	Neural Network	Logistic Regression
Gender	**71.53(0.37)**	66.73(1.04)	66.77(1.41)
Age	**43.09(1.56)**	41.74(0.71)	43.01(1.63)
Marital Status	**63.94(1.85)**	60.81(2.54)	62.00(1.14)
Residence	29.98(2.09)	36.41(1.49)	**36.78(0.91)**
Job	26.08(0.75)	**30.32(1.54)**	29.92(0.50)

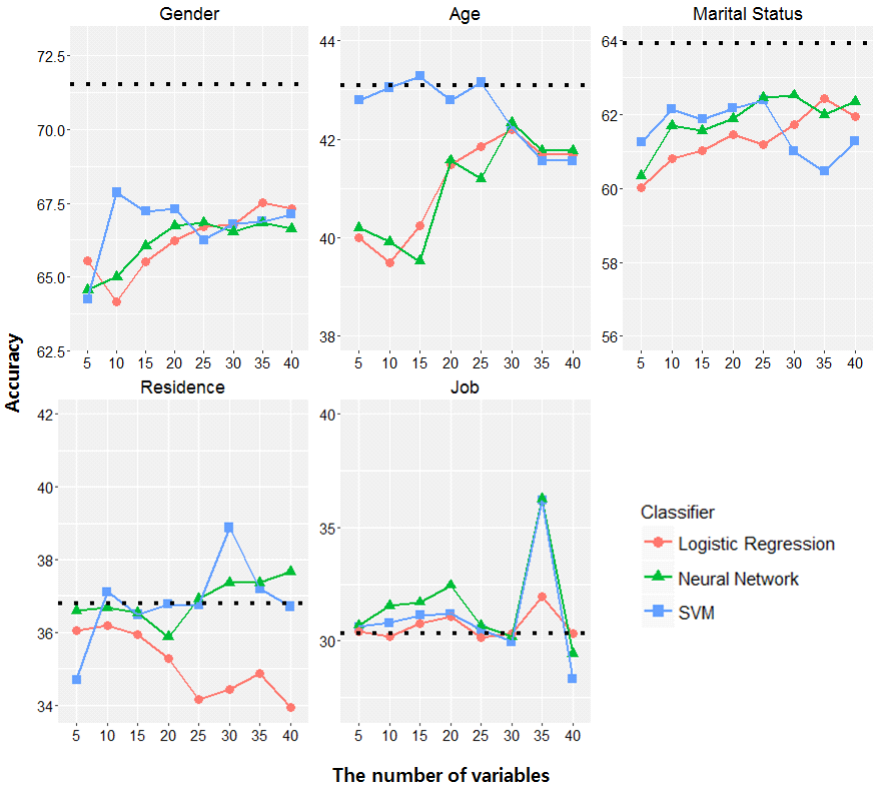
* Figures in parentheses refer to the standard deviation and all figures shows percentage.

사용자의 인구통계특성 중 성별, 연령, 혼인여부는 SVM을 적용하여 예측 모델을 학습한 결과, 정확도가 각각 71.53%, 43.09%, 63.94%로 가장 높음을 알 수 있다. 거주지는 로지스틱 회귀모형을 적용했을 때 정확도가 36.78%로 가장 높고, 직업은 신경망 모델을 적용했을 때 정확도가 30.32%로 가장 높다. 종속변수로 사용된 성별, 연령, 혼인여부, 거주지, 직업은 각각 2개, 5개, 2개, 13개, 20개의 클래스를 가진다. <Table 3>의 각 종속변수에 대한 예측 모델의 최고 정확도는 아래의 변수축소 방법들을 적용하여 예측 모델 생성 시 성능 비교를 위한 기준이 된다.

4.3.2 의사결정나무를 이용한 변수축소 방법 적용

본 실험에서는 변수축소를 위한 방법 중 하나로 의사결정나무 기법의 알고리즘 중 하나인 C5.0을 사용하였고 출력 유형으로 부스팅 10회, 교차 타당성 10회를 적용하였다. C5.0을 이용하여 종속변수로 사용된 인터넷 사용자의 인구통계특성에 대해 각각 예측 변수 중요도를 측정한 후 중요도가 높은 변수들을 5개에서 40개까지 5단위로 늘려가며 예측 모델의 설명변수로 사용하다. 각각의 인구통계특성에 대한 예측 모델의 평가 결과는 <Figure 3>과 같다. 각 그림에 있는 점선(dotted line)은 각각의 분류 기법에 모든 설명변수를 사용하여 예측 모델을 생성 했을 때 가장 높은 정확도를 나타낸 선이다.

실험 결과 각 종속변수의 클래스 개수에 따라 예측 모델의 정확도에 차이가 존재함을 알 수 있다. 먼저 이분형 변수인 성별과 혼인여부 그리고 다분형 변수이지만 상대적으로 클래스의 수가 적은 연령은 의사결정나무를 이용한 변수축소 방법을 적용한 예측 모델의 정확도가 모든 설명변수를 적용한 예측 모델의 정확도보다 낮거나 미비한 차이가 존재한다. 클래스의 수가 상대적으로 많은 거주지와 직업에 대한 실험 결과 모든 설명변수를 사용했을 때보다 의사결정나무를 이용한 변수축소 방법을 적용하여 설명변수가 각각 30개, 35개 일 때 예측모델의 정확도가 가장 높은 것을 알 수 있다. 이는 클래스가 많은 다분형 변수는 모든 설명변수를 사용하여 예측 모델



〈Figure 3〉 Accuracy for dimension reduction based on decision tree

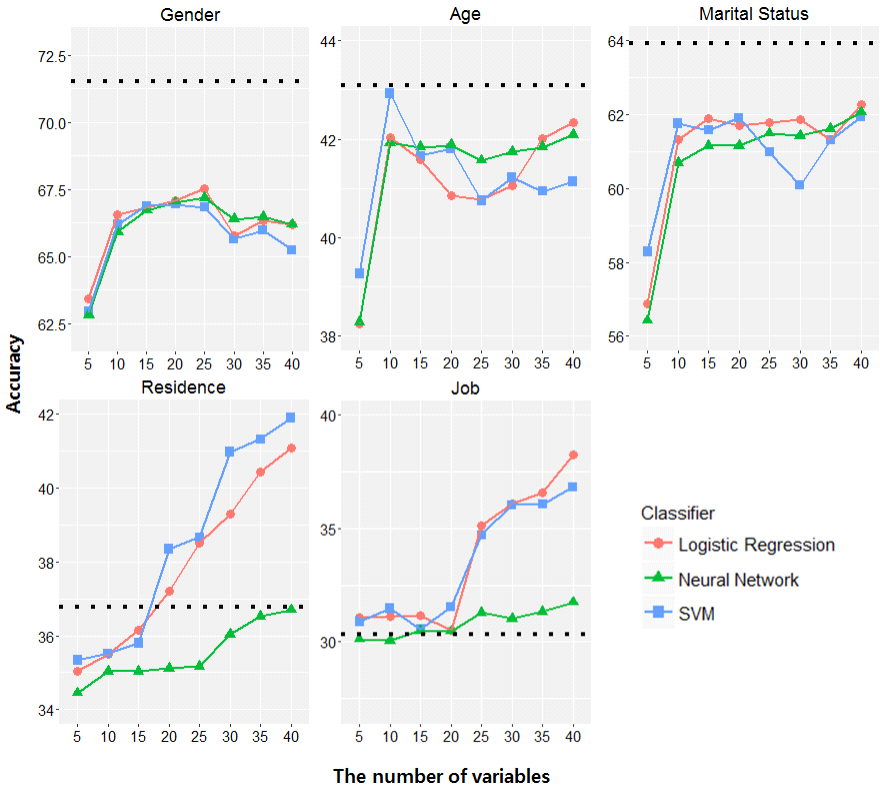
학습 시 과적합 문제가 발생하며 의사결정나무를 이용한 변수축소 방법은 클래스가 많은 다분형 종속변수에 대한 예측 모델의 성능을 높이는데 효과가 있음을 의미한다.

4.3.3 주성분분석을 이용한 변수축소 방법 적용

본 연구에서는 주성분분석은 변수축소를 위한 방법으로 사용되었으며 주성분분석을 위해 상관계수 행렬을 사용하였다. 주성분분석을 이용한 변수축소 방법은 주성분의 개수를 5개에서 40개까지 5단위로 늘려가며 예측 모델의 설명변수로 사용하였다. 추출된 주성분을 이용하여 각각의 종속변수를 예측한 결과는 <Figure 4>와 같다.

각 그림에 있는 점선(dotted line)은 각각의 분류 기법에 모든 설명변수를 사용하여 예측 모델을 생성 했을 때 가장 높은 정확도를 나타낸 선이다.

이분형 종속변수인 성별과 혼인여부 그리고 다분형 종속변수이지만 클래스의 개수가 비교적 적은 연령은 주성분을 이용한 변수축소 방법을 적용한 예측 모델의 정확도보다 모든 설명변수를 사용한 예측 모델의 정확도가 높은 것을 알 수 있다. 다분형 종속변수인 거주지와 직업에 대한 예측 모델은 설명변수로 사용된 주성분의 개



〈Figure 4〉 Accuracy for dimension reduction based on PCA

수가 25개 이상일 때 모든 설명변수를 사용한 예측 모델의 정확도보다 높아짐을 알 수 있다. 이는 의사결정나무를 이용한 변수축소 방법과 같은 맥락으로 주성분분석을 이용한 변수축소 방법도 다분형 변수 종속변수에 대한 예측 모델의 성능을 높이는데 효과가 있음을 의미한다.

4.3.4 군집분석을 활용한 변수축소 방법 적용

마지막으로 변수축소 방법을 적용하기 위해 인터넷 사용자를 기준으로 웹사이트 카테고리, 시간대, 요일, 월, 뉴스사이트 카테고리 페이지뷰 비율에 대해 각각 군집분석을 시행하였다. 실험에 사용된 군집분석 알고리즘은 k-평균(k-means)이며 군집 개수를 결정하기 위해 각 객체가 적절한 군집에 배치되었는지를 측정하는 평균 실루엣(silhouette)을 참고하였다(Kaufman and Rousseeuw, 2009).

<Table 4>는 웹사이트 카테고리별 페이지뷰 비율 변수 22개, 시간대별 페이지뷰 비율 변수 4개, 요일별 페이지뷰 비율 변수 7개, 월별 페이지뷰 비율 변수 12개, 뉴스사이트 카테고리별 페이지뷰 비율 변수 12개에 대해 각각 군집분석을 적용한 결과를 나타낸 것이다. 구체적으로 <Table 4>를 통해 웹사이트 카테고리별 페이지뷰 비율

〈Table 4〉 Result of Clustering

Variable	no	Summary for cluster
Ratio of pageviews for website category(22)	1	Internet and Computer type
	2	Education and Health type
	3	Shopping type
	4	Community type
	5	Entertainment type
Ratio of pageviews for time(4)	1	Day type
	2	Night type
Ratio of pageviews for day(7)	1	Week type
	2	Weekend type
Ratio of pageviews for month(12)	1	Summer/Fall type
	2	Spring/Winter type

Variable	no	Summary for cluster
Ratio of pageviews for news website category (12)	1	Sports news type
	2	world news type
	3	IT·Science news type
	4	Column news type
	5	Preference that not news
	6	Life and Culture news type
	7	Community news type
	8	Hankyoreh news type
	9	Entertainments news type
	10	Economic and Political news type

* The figures in parentheses are the number of categories for each variable.

변수는 5개, 시간대별 페이지뷰 비율 변수는 2개, 요일별 페이지뷰 비율 변수는 2개, 월별 페이지뷰 비율 변수 2개, 뉴스사이트 카테고리별 페이지뷰 비율 변수는 10개의 군집으로 구분 가능함을 알 수 있다.

군집분석을 적용하여 57개의 연속형 변수를 축소시킨 5개의 범주형 변수와 군집분석을 적용하지 않은 7개의 연속형 변수를 예측 모델의 설명변수로 사용하였다. 이를 사용한 각각의 인구통계특성 예측에 대한 실험 결과는 <Table 5>와 같다. <Table 5>의 굵은 수치는 모든 설명변수를 사용한 예측 모델보다 군집분석을 활용한 변수축소 방법을 적용한 예측 모델의 정확도가 높은 것이다. 각각의 종속변수에 대해 모든 설명변수를 사용한 예측 모델의 정확도보다 군집분석을 활용한 변수축소 방법을 적용한 예측 모델의 정확도가 대부분 낮음을 알 수 있다. 이는 군집분석을 활용한 변수축소 방법은 여러 개의 연속형 변수를 하나의 범주형 변수로 축소시켰기 때문에 데이터의 정보 손실율이 높았기 때문인 것으로 판단된다.

〈Table 5〉 Accuracy for dimension reduction based on clustering

	SVM	Neural Network	Logistic Regression
Gender	63.13(2.04)	64.02(1.96)	64.34(1.91)
Age	34.67(1.56)	39.19(0.86)	40.36(1.35)
Marital Status	54.34(1.87)	60.90(0.49)	59.67(1.85)
Residence	26.45(2.01)	36.87(0.83)	**37.12(0.87)**
Job	19.97(6.87)	30.20(0.51)	29.82(0.14)

* Figures in parentheses refer to the standard deviation and all figures shows percentage.

4.4 결과 요약 및 논의

본 연구에서는 예측 모델 생성 시 발생하는 차원의 저주 및 과적합 문제를 해결하기 위해 총 3가지 관점의 변수축소 방법을 제안하였다. 각각의 종속변수와 변수축소 방법에 따른 예측 모델의 최대 정확도는 <Table 6>과 같다. 이분형 종속변수인 성별과 혼인여부의 경우 변수축소 방법을 사용하지 않고 모든 설명변수를 사용한 예측 모델의 정확도가 가장 높음을 알 수 있으며

이때 사용된 분류 기법은 SVM이다. 반면에, 비교적 적은 수의 클래스를 가진 다분형 종속변수인 연령에 대한 예측 모델은 의사결정나무를 이용한 변수축소 방법을 적용하고 신경망 분류 기법을 이용한 예측 모델의 정확도가 가장 높으며 이때 예측 변수 중요도가 높은 상위 15개 변수를 설명변수로 사용하였다.

또한 각각 13개, 20개로 많은 수의 클래스를 가지는 다분형 종속변수인 거주지와 직업에 대한 예측 모델은 주성분분석을 이용한 변수축소 방법을 적용하고 신경망 분류 기법을 이용한 예측 모델의 정확도가 가장 높음을 알 수 있다. 구체적으로 거주지의 경우 주성분분석을 이용한 변수축소 방법을 적용하고 신경망 분류 기법을 이용하였을 때 예측 모델의 정확도가 41.90%로 가장 높음을 알 수 있다. 이 경우 사용된 주성분의 개수는 40개이다. 마찬가지로 직업의 경우에도 주성분분석을 이용한 변수축소 방법을 적용하고 신경망 분류 기법을 이용하였을 때 예측 모델의 정확도가 38.25%로 가장 높은 것을 확인하였으며 이 경우 사용된 주성분 개수는 40개이다.

〈Table 6〉 Summary of experiments

	Accuracy (%)	classifier	Method of variable reduction
Gender	71.53	SVM	None (Using all variables)
Age	42.38	Neural Network	decision tree
Marital Status	63.94	SVM	None (Using all variables)
Residence	41.90	Neural Network	PCA
Job	38.25	Neural Network	PCA

본 연구는 모든 설명변수를 사용하여 이분형 종속변수에 대한 예측 모델 생성 시 가장 적합한 분류 기법은 SVM 분류 기법이며 변수축소 방법을 적용하여 다분형 종속변수에 대한 예측 모델 생성 시 가장 적합한 분류 기법은 신경망 분류 기법으로 확인되었다. 또한, 본 연구에서는 클래스의 수가 많은 다분형 종속변수에 대해 의사결정나무나 주성분분석을 이용한 변수축소 방법을 적용하여 가장 높은 정확도를 가지는 예측 모델을 생성하였다. 그리고 종속변수의 클래스의 수가 많아질수록 변수축소 방법을 적용하는 것이 예측 모델의 정확도를 높이는 방법임을 확인하였다. 반면, 각 종속변수에 대해 모든 설명변수를 사용한 예측 모델보다 군집분석을 활용한 변수축소 방법을 적용한 예측 모델이 낮은 정확도를 보였다. 이는 군집분석을 활용한 변수축소 방법은 여러 개의 연속형 변수를 하나의 범주형 변수로 축소시켰기 때문에 데이터의 정보 손실율이 높았으며 이로 인해 예측 모델이 훈련용 데이터에 부적합(underfitting)되었기 때문이라 판단된다. 따라서 무조건적인 변수축소는 예측 모델의 성능을 개선시키는 방법이 아니지만 일정 수준 이상의 클래스를 가지는 다분형 종속변수에 대해 변수축소 방법을 적용하는 것은 예측 모델의 차원의 저주 및 과적합의 문제를 해결할 수 있으며 예측 모델의 정확도와 신뢰성을 높일 수 있음을 확인하였다.

5. 결론

5.1 연구 요약

본 연구는 클릭스트림 데이터(clickstream data)

에서 추출한 온라인 행위 정보를 이용한 인터넷 사용자의 인구통계특성 예측에 관한 연구로 실험을 위해 클릭스트림 데이터를 활용하여 인터넷 사용자의 온라인 행위 정보와 인구통계특성 정보를 포함하는 사용자 프로파일(profile)을 생성하였다.

온라인 행위 정보는 크게 선호사이트, 사용 패턴, 검색행위, 관심사로 구분된다. 선호사이트에 대한 변수로 웹사이트의 특성에 따라 분류된 22개의 카테고리별 페이지뷰 비율, 카테고리에 대한 변동계수가 있고 사용 패턴에 대한 변수로 총 페이지뷰, 총 방문일수, 시간대별 페이지뷰 비율, 요일별 페이지뷰 비율, 월별 페이지뷰 비율, 시간에 대한 변동계수, 요일에 대한 변동계수, 월에 대한 변동계수가 있다. 그리고 검색행위는 총 검색키워드 수로 설명되며 관심사는 뉴스사이트 카테고리별 페이지뷰 비율로 설명된다.

먼저 사용자의 온라인 행위 정보를 나타내는 64개의 설명변수로 SVM, 신경망, 로지스틱회귀분류 기법을 이용하여 각각의 인구통계특성에 대한 예측 모델을 생성하였다. 하지만 온라인 행위 정보를 나타내는 모든 설명변수를 사용하여 예측 모델을 학습할 경우 데이터 차원이 증가하여 모델의 정확도를 유지하기 위해 필요한 데이터의 수가 기하급수적으로 증가하는 차원의 저주에 대한 문제가 발생 할 수 있다. 또한 불필요한 변수의 사용으로 인해 모델의 과적합 문제가 발생할 가능성이 높다. 따라서 본 연구에서는 (1)의사결정나무를 이용한 변수축소, (2)주성분분석을 이용한 변수축소, (3)군집분석을 활용한 변수축소 방법들을 제안하고 이에 대한 모델 성능을 평가하기 위해 정확도를 이용하였다. 그 결과 클래스의 수가 많은 다분형 종속변수에 대한 예측 모델은 의사결정나무와 주성분분석을 이용한 변수축소 방법을 적용하였을 때 예측 모델의 정확도가 가장 높았다. 반면, 군집분석을 활용한 변수축소 방법을 적용한 예측 모델의 정확도는 모든 설명변수를 사용한 예측 모델의 정확도보다 낮았다. 이는 무조건적인 변수축소 방법은 예측 모델의 성능 향상에 영향을 미치지 않음을 의미한다.

5.2 시사점 및 의의

본 연구는 이론적인 측면에서 2가지 의의가 있다. 첫째, 클릭스트림 데이터를 활용하여 인터넷 사용자의 인구통계특성을 예측할 수 있다는 것이다. 다시 말해 익명의 온라인 활동이 점자 증가하고 있는 현재시점에서 인터넷 사용자의 온라인 활동 기록만으로도 인구통계특성을 예측할 수 있다는 점에서 의의가 있다. 둘째, 예측 모델 생성 시 발생할 수 있는 차원의 저주 및 과적합 문제를 해결할 수 있는 여러 가지 변수축소 방법들을 제안하였다. 이를 통해 어떤 종속변수에 대해 어느 변수축소 방법을 사용하는 것이 효과적인지를 제시하였다는 점에서 의의가 있다. 향후 다양한 변수축소 방법 및 분류 기법을 활용한 예측 모델 성능 개선 연구에 본 연구에서 제안한 변수축소 방법을 결합하여 보다 높은 정확도를 가지는 예측 모델을 생성할 수 있을 것이다. 이러한 이론적 의의들을 토대로 익명의 인터넷 사용자에 대한 인구통계특성은 클릭스트림 데이터를 통해 확인 가능하며 이를 타겟 마케팅, 개인화된 광고 등에 활용 가능하다는 점에서 실무적 시사점이 있다.

5.3 한계점 및 향후 연구 방향

본 연구는 클릭스트림 데이터를 이용하여 사

용자의 인구통계특성을 예측하였고 예측 모델 생성 시 발생하는 차원의 저주 및 과적합 문제를 해결하기 위해 총 3가지 관점의 변수축소 방법을 제시하였다. 하지만 3가지 관점의 방법론 중 군집분석을 활용한 변수축소 방법은 예측 모델의 정확도 개선에 도움이 되지 않았다. 이는 여러 개의 연속형 변수를 하나의 범주형 변수로 축소시켰기 때문에 정보의 손실율이 커져 예측 모델의 성능이 개선되지 않은 것으로 판단된다. 따라서 향후 연구에서는 군집분석을 활용하여 연속형 변수를 범주형 변수로 축소시킴에 있어 어떤 변수에 대한 변수축소가 예측 모델의 성능을 저하시키는지, 몇 개의 연속형 변수를 범주형 변수로 축소시켜야 예측 모델의 성능이 개선되는지에 대한 확장 연구가 가능하다. 또한 본 연구에서는 의사결정나무와 주성분분석을 이용한 변수축소 방법을 적용함에 있어 예측 모델 학습 시 사용된 설명변수의 개수를 5단위로 변화시켜가며 예측 모델의 정확도를 측정하였다. 이에 향후 연구에서는 더 세분화된 단위의 사용하여 예측 모델의 정확도 개선에 대한 구체적인 기준점을 찾을 수 있을 것이라 기대된다.

참고문헌(References)

Ban, H. and Y. Kwon, “The Study of the Usage Correlation between Portal and Traditional News Media”, *Korean Journal of Journalism & Communication Studies*, Vol.51, No.1 (2007), 399~426.

Banlioni and Miriam, et al., “Preprocessing and mining web log data for web personalization”, *Congress of the Italian Association for Artificial Intelligence*. Springer Berlin Heidelberg, 2003, 237~249.

Boutsidis, Christos, M. W. Mahoney and P. Drineas, “Unsupervised feature selection for principal components analysis”, *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, 61~69.

Bucklin. Randolph E et al., “Choice and the Internet: From clickstream to research stream”, *Marketing Letters*(2002), 245~258.

Cho, K. and H. Park, “A study on 3-step complex data mining in society indicator survey”, *Journal of the Korean Data & Information Science Society*, Vol.23, No.5(2012), 983~992.

Choi, S., Y. Hyun and N. Kim, “Improving Performance of Recommendation Systems Using Topic Modeling”, *Journal of Intelligence and Information Systems*, Vol.22, No.1(2015), 77~93.

De Bock, W. Koen and V. D. Poel, “Predicting website audience demographics for web advertising targeting using multi-website clickstream data”, *Fundamenta Informaticae*, Vol.98, No.1(2010), 49~70.

Eleonora Ivanova, “Predicting website audience demographics based on browsing history”, Master’s Thesis, Information and Service Management, Aalto University, 2013.

Foody, M. Giles and A. Mathur, “A relative evaluation of multiclass image classification by support vector machines”, *IEEE, Transactions on geoscience and remote sensing*, Vol.42, No.6(2004), 1335~1343.

Gallagher, K. and J. Parsons, “A framework for targeting banner advertising on the Internet”, *Proc. 30th Hawaii International Conference on System Sciences(HICSS 30)*, 1997.

Goel, Sharad, M. Jake, Hofman and M. I. Sirer, "Who Does What on the Web: A Large-Scale Study of Browsing Behavior." *ICWSM*, 2012.

Han, S et al. "Real-Time Purchase Probability Prediction Using Clickstream Data of Internet Storefronts", *Entrue Journal of Information Technology*, Vol.11, No.1(2012), 101~110.

Huang, Zan et al., "Credit rating analysis with support vector machines and neural networks: a market comparative study", *Decision support systems*, Vol.37, No.4(2004), 543~558.

Jones et al., "I know what you did last summer: query logs and user privacy", *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. ACM*, 2007, 909~914.

Kaufman, Leonard, J. Peter and Rousseeuw, "Finding groups in data: an introduction to cluster analysis", Vol. 344. John Wiley & Sons, 2009.

Kim, I., "Predicting audience demographics of web sites using local cues", Doctoral dissertation, David Eccles School of Business, The University of Utah, 2011.

Kim, T. and H. Ahn, "A Hybrid Under-sampling Approach for Better Bankruptcy Prediction", *Journal of Intelligence and Information Systems*, Vol.21, No.2(2015), 173~190.

Kim, Y. et al., "A Study on Method for User Gender Prediction Using Multi-Modal Smart Device Log Data", *The Journal of Society for e-Business Studies,* Vol.21, No.1(2016), 147~163.

Lee, K. and H. Lee, "A Study on the Combined Decision Tree(C4.5) and Neural Network Algorithm for Classification of Mobile Telecommunication Customer", *Journal of Intelligence and Information Systems*, Vol.9, No. 1(2003).

Moe and W. Wendy, "Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream", *Journal of consumer psychology*, Vol.13, No.1(2003), 29~39.

Montgomery, A. L et al., "Modeling online browsing and path analysis using clickstream data", *Marketing Science,* Vol.23, No.4 (2004), 579~595.

Murray, D. and K. Durrell, "Inferring demographic attributes of anonymous internet users", Web usage Analysis and User Profiling Workshop, Springer, 2000, 7~20.

Park, Y. -H. and S. F. Peter, "Modeling browsing behavior at multiple websites", *Marketing Science*, Vol.23, No.3(2004), 280~303.

Poindexter, M. Paula and E. M. Maxwell, "Revisiting the civic duty to keep informed in the new media environment", *Journalism & Mass Communication Quarterly*, Vol.78, No.1(2001), 113~126.

Provost, Foster and T. Fawcett, "Data Science for Business: What you need to know about data mining and data-analytic thinking", O'Reilly Media, Inc., 2013.

Rumelhart, David E., E. Geoffrey, Hinton and R. J. Williams, "Learning internal representations by error propagation", No. ICS-8506. CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE, 1985.

Abstract

Clickstream Big Data Mining for Demographics based Digital Marketing

Jiae Park*·Yoonho Cho**

The demographics of Internet users are the most basic and important sources for target marketing or personalized advertisements on the digital marketing channels which include email, mobile, and social media. However, it gradually has become difficult to collect the demographics of Internet users because their activities are anonymous in many cases. Although the marketing department is able to get the demographics using online or offline surveys, these approaches are very expensive, long processes, and likely to include false statements. Clickstream data is the recording an Internet user leaves behind while visiting websites. As the user clicks anywhere in the webpage, the activity is logged in semi-structured website log files. Such data allows us to see what pages users visited, how long they stayed there, how often they visited, when they usually visited, which site they prefer, what keywords they used to find the site, whether they purchased any, and so forth. For such a reason, some researchers tried to guess the demographics of Internet users by using their clickstream data. They derived various independent variables likely to be correlated to the demographics. The variables include search keyword, frequency and intensity for time, day and month, variety of websites visited, text information for web pages visited, etc. The demographic attributes to predict are also diverse according to the paper, and cover gender, age, job, location, income, education, marital status, presence of children. A variety of data mining methods, such as LSA, SVM, decision tree, neural network, logistic regression, and k-nearest neighbors, were used for prediction model building. However, this research has not yet identified which data mining method is appropriate to predict each demographic variable. Moreover, it is required to review independent variables studied so far and combine them as needed, and evaluate them for building the best prediction model. The objective of this study is to choose clickstream attributes mostly likely to be correlated to the demographics from the results of previous research, and then to identify which data mining method is fitting to predict

* Dept. of Data Science, Kookmin University
** Corresponding Author: Yoonho Cho
School of Business Administration, Kookmin University
77 Jeongneung-ro, Seongbuk-gu, Seoul 02707, Korea
Tel: +82-2-910-4950, Fax: +82-2-910-5209, E-mail: www@kookmin.ac.kr

each demographic attribute. Among the demographic attributes, this paper focus on predicting gender, age, marital status, residence, and job. And from the results of previous research, 64 clickstream attributes are applied to predict the demographic attributes. The overall process of predictive model building is compose of 4 steps. In the first step, we create user profiles which include 64 clickstream attributes and 5 demographic attributes. The second step performs the dimension reduction of clickstream variables to solve the curse of dimensionality and overfitting problem. We utilize three approaches which are based on decision tree, PCA, and cluster analysis. We build alternative predictive models for each demographic variable in the third step. SVM, neural network, and logistic regression are used for modeling. The last step evaluates the alternative models in view of model accuracy and selects the best model. For the experiments, we used clickstream data which represents 5 demographics and 16,962,705 online activities for 5,000 Internet users. IBM SPSS Modeler 17.0 was used for our prediction process, and the 5-fold cross validation was conducted to enhance the reliability of our experiments. As the experimental results, we can verify that there are a specific data mining method well-suited for each demographic variable. For example, age prediction is best performed when using the decision tree based dimension reduction and neural network whereas the prediction of gender and marital status is the most accurate by applying SVM without dimension reduction. We conclude that the online behaviors of the Internet users, captured from the clickstream data analysis, could be well used to predict their demographics, thereby being utilized to the digital marketing.

Key Words : Big Data, Clickstream Data, Demographics, Online Behavior, Classification, Variable Reduction, Accuracy

Received : August 17, 2016 Revised : September 19, 2016 Accepted : September 24, 2016
Publication Type : Regular Paper Corresponding Author : Yoonho Cho

저 자 소 개

박 지 애
현재 국민대학교 데이터사이언스학과 석사과정에 재학 중이며, 명지대학교 수학과에서 학사학위를 취득하였다. 주요 관심분야는 데이터마이닝, 고객관계관리, 마케팅 애널리틱스, 기상데이터분석 등이다.

조 윤 호
현재 국민대학교 경영학부 빅데이터경영통계전공 교수로 재직 중이다. 서울대학교 계산통계학과를 졸업하고, KAIST 경영정보공학과에서 석사, KAIST 경영공학과에서 박사학위를 취득하였으며, LG전자(주)에서 6년간 주임연구원으로 재직하였다. 주 연구분야는 비즈니스애널리틱스, 빅데이터 마이닝, 추천시스템, 소셜네트워크분석, 고객관계관리 등이다.