JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 19, No. 10 : 2232~2240 Oct. 2015

빅데이터 환경에서 사용자 거래 성향분석을 위한 머신러닝 응용 기법

최도현[1] · 박중오[2*]

The Application Method of Machine Learning for Analyzing User Transaction Tendency in Big Data environments

Do-hyeon Choi[1] · Jung-oh Park[2*]

[1]Department of Computer Science, Soongsil University, Seoul 156-743, Korea
[2*]Information & Communication Engineering, Dongyang Mirae University, Seoul 152-714, Korea

요 약

최근 빅데이터 분야에서는 고객의 흥미가 높은 상품이나 과거 구매 내역 등 기존 보유한 데이터를 수집 및 재가공하여 사용자의 거래성향을 분석(상품 추천, 판매 예측 등)하는데 활용하려는 추세이다. 기존 사용자의 성향 관련 연구들은 조사시기와 대상의 범위가 한정적이며 세부 상품에 대한 예측이 어렵고, 실시간성이 없기 때문에 트렌드에 적절한 빠른 판매 전략을 도입하기가 어려운 단점이 존재한다. 본 논문은 기계학습 알고리즘 응용하여 사용자의 거래성향 분석에 활용한다. 기계학습 알고리즘 응용 결과 세부 상품별 추론할 수 있는 다양한 지표를 추출할 수 있음을 증명하였다.

ABSTRACT

Recently in the field of Big Data, there is a trend of collecting and reprocessing the existing data such as products having high interest of customers and past purchase details to be utilized for the analysis of transaction propensity of users(product recommendations, sales forecasts, etc). Studies related to the propensity of previous users has limitations on its range of subjects and investigation timing and difficult to make predictions on detailed products with lack of real-time thus there exists difficult disadvantages of introducing appropriate and quick sales strategy against the trend. This paper utilizes the machine learning algorithm application to analyze the transaction propensity of users. As a result of applying the machine learning algorithm, it has demonstrated that various indicators which can be deduced by detailed product were able to be extracted.

키워드 : 빅데이터, 머신 러닝, 딥 러닝, 성향 분석, 데이터마이닝

Key word : Big data, Machine Learning, Deep Learning, Tendency Analysis, Data Mining

Received 31 August 2015, **Revised** 23 September 2015, **Accepted** 05 October 2015

*** Corresponding Author** Jung-oh Park(E-mail:jopark13@dongyang.ac.kr, Tel:+82-2-2610-5169)
Information & Communication Engineering, Dongyang Mirae University, Seoul 152-714, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2015.19.10.2232 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

초고속 인터넷 도입 이후 2000년대 이후 급격한 인터넷 사용자 증가와 온라인 마켓 활성화에 따라 각 온라인 마켓 업체들은 효율적인 마케팅을 위해 사용자의 구매성향을 분석하려는 다양한 분서기법과 연구들이 제안되었다[1]. 기존 데이터마이닝(Data Mining) 분야에서는 성향 분석을 위한 상품 종류 및 품질, 웹 사이트 및 사용자의 특성, 성별이나 연령대, 서비스 만족도 등 특정 데이터를 추출하기 위해서 다양한 연구들이 제안되었지만 결론적으로 데이터 분석에 직접 큰 비용과 시간을 소비해야 했다[2].

최근 모바일, 클라우드, 빅데이터 등 IT 환경 트렌드 변화에 따라 온라인 마켓에서는 상품추천을 할 수 있는 방법으로 인공지능의 한 영역인 기계학습이 재 주목받고 있다[3]. 기존 기계학습 분야는 IT 업계에서 스팸 필터(Spam Filter), 얼굴 인식, 일기 예보, 상품 추천, 장비 고장 예측 등 서비스업, 제조업 등 특정 분야에서 활용되어 왔다[4].

빅데이터 환경을 활용한 기계학습은 기존 분야에서 축척된 정형화된 데이터(구조적, 수치로 표현된 데이터)이외에 비정형 데이터까지 분석 범위를 확장하였고, 병렬처리 및 분산처리 파일 시스템의 장점으로 기존 기계학습 알고리즘의 연산 비효율성과 저장 공간의 문제를 해결하였다[5].

본 논문에서는 빅데이터 환경에서 최근 활용되는 기계학습 알고리즘을 응용하여 기존의 분석과 판단이 어려웠던 사용자의 상품 거래에 대한 성향을 분석한다. 2장은 관련연구, 3장은 제안 기법, 4장은 성능 평가, 5장 결론으로 마친다.

Ⅱ. 관련 연구

기계학습이란 수집된 다양한 데이터 분석을 할 수 있는 기준(알고리즘)을 가지고 학습을 통해 주어진 일에 대한 해결책 제시를 자동화하는 것을 의미한다. 기존 정체되었던 기계학습과 관련된 연구들은 최근 등장한 빅데이터 기술 분야에서 점점 실현 가능성이 커지고 있다[6]. 본 관련연구에서는 본 논문에서 활용한 기계학습 알고리즘에 대하여 설명한다.

2.1. Naïve Bayes Algorithm

나이브 베이스는 특정 단어에 대한 문서 분류와 횟수를 기록하는데 활용되며 베이스 정리의 일부분으로 두 개의 분류항목간에 높은 확률을 가지는 항목을 선택하는 방법이다[7]. 베이스 규칙(Bayes'rule)에서는 조건부 확률(Conditional probability)로 데이터를 분류하기 위한 방법으로 다음과 같은 공식을 사용한다[7].

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)}$$

데이터 분류에 x를 입력 데이터, c를 분류 항목으로 정의했을 때 $p(x|c)$를 알고 있는 상태에서 $p(c|x)$를 알 수 있다. 조건부 확률로 분류하기 위해서는 다음과 같은 두 가지 확률을 찾을 수 있다. $p1(x,y) \rightarrow p2(x,y)$이면 분류 항목 1, $p1(x,y) \rightarrow p2(x,y)$이면 분류 항목 2에 속한다. 위 2가지 규칙은 다음과 같은 베이스 규칙으로 적용할 수 있다[8].

$$p(c_i|x,y) = \frac{p(x,y|c_i)p(c_i)}{p(x,y)}$$

두 가지 규칙 $p(c1|x,y)$와 $p(c2|x,y)$를 비교하여 x, y로 확인된 데이터를 기준으로 분류항목 $c1$에 속할 확률을 구한다. 이 경우 $p(c1|x,y) \rightarrow p(c2|x,y)$이면 $c1$, $p(c1|x,y) \leftarrow p(c2|x,y)$이면 분류 항목 $c2$에 속한다. 위와 같은 베이스 규칙을 사용하여 알려진 데이터 개수 i로부터 알려지지 않은 것을 계산할 수 있다.

2.2. Decision Tree Algorithm

의사결정 트리는 DB마케팅, CRM, 시장조사, 광고조사, 의학연구, 품질관리 등 분류 기술 중 가장 일반적으로 사용되는 방법으로 데이터 집합을 분류하는데 사용할 수 있다. 최초 부모 노드로부터 데이터 집합을 분류될 때 까지 재귀적인 분류절차를 반복하는 형태를 가지고 있다[9]. 데이터 분할을 결정하는 방법 중 양자화(Quantitative)방법을 적용하며, 정보이득(Information Gain)을 계산하여 각 노드별 속성에 대해 확인하여 분할을 진행한다. 다음은 분류항목이 선택될 확률 계산하기 위해 다음과 같은 공식을 사용한다[10].

$$L = \log_2 p(x_i)$$

여기서 $p(x_i)$는 x_i라는 분류 항목이 선택될 확률이다. 정보이득에 대한 정보 측정 방법은 다음과 같은 엔트로피(Entropie) 계산 공식을 사용한다[10].

$$H = -\sum_{i=1}^{n}(x_i)\log_2 p(x_i)$$

분류된 데이터 항목 L의 확률(x_i)을 이용하여 엔트로피 H를 계산하여 모두 더한다. 마지막으로 비교하기 전, 후 변화를 비교하여 가장 좋은 속성의 색인을 반환하고, 앞 과정을 반복하여 정보이득(엔트로피)이 가장 높은 속성에 대해서 데이터 분류한다.

2.3. Knn(K-nearest neighbor) Algorithm

K-최근접 이웃은 얼굴인식, 핑커프린트 등 패턴인식 분야에서 가장 잘 알려진 알고리즘으로 데이터 분류를 위해 기존의 모든 데이터(분류된 데이터 항목 필요)와 새로운 데이터를 비교하여 상위 k개의 가장 유사한 데이터와의 거리를 측정하여 데이터를 분류하는 방법이다[11]. 거리를 측정하는 방법은 유클리드 거리(Euclidean distance)를 사용하며 공식은 다음과 같은 공식을 사용한다[11].

$$d = \sqrt{(xA_0 - xB_0)^2 + (xA_1 - xB_1)^2}$$

두 가지 데이터 그룹 xA와 xB간의 속성 값을 가지고 거리를 계산할 수 있다. 각 속성 값의 거리를 구하기 위해서는 정수형 수치 값을 속성 값으로 사용하며 이를 위해 데이터의 정규화가 필요하다. 정규화는 0에서 1또는 -1에서 1 등 사용자가 정의할 수 있으며 다음과 같은 공식을 사용한다[12].

$$V = (oldV - \min V) / (\max V - \min V)$$

정규화 V는 집합 내 가장 작은 값 $\min V$와 가장 큰 값 $\max V$를 사용하여 정규화를 진행한다. 정규화 과정에서 사용된 최대/최소값은 가까운 거리를 판단하여 분류되지 않은 유형에 대한 거리를 예측하는데 사용할 수 있다.

2.4. Apriori Algorithm

Apriori 알고리즘은 연관규칙, 선호도, 정보여과 등 데이터 변수들에서 관찰 되는 주요한 관계를 가장 적합하게 설명할 수 있는 규칙을 찾는 알고리즘으로 높은 다차원이나 복잡한 관계를 가지는 데이터 간에 중요 연관성을 찾는데 사용될 수 있다[13].

Apriori 알고리즘에서 관계는 빈발 아이템 집합(Frequent item sets) 또는 연관규칙(Association rules) 두 가지 형태로 표현한다. 빈발 아이템 집합은 함께 자주 발생하는 아이템들을 모은 것이며, 연관규칙은 아이템 간의 관계에 강도가 존재한다고 제안하는 것이다.

특정 데이터 집합에서 위 두 가지 방법을 사용하여 아이템 집합에 대한 관계 여부를 판단할 수 있다. 특정 아이템 A에서 B에 대한 신뢰도를 지지도는 다음과 같은 공식을 사용한다[14].

$$G = \frac{DG_i}{DG_{total}} \quad {}^{IF\{A\}\rightarrow\{B\}} S = G(\{A,B\}) / G(\{A\})$$

지지도는 데이터 그룹에 특정 데이터가 포함된 데이터 집합의 비율 G와 신뢰도 S는 연관규칙으로 정의되어 연관성이 많은 데이터들을 그룹화 하는 군집화의 일종으로, 목적을 동시에 만족하는 가능성이 큰 데이터들을 찾는데 사용할 수 있다[14].

Ⅲ. 제안 기법

본 논문에서 사용한 기계학습 알고리즘은 Naïve Bayes, Decision Tree, Knn(K-nearest neighbor), Apriori를 상호 응용하였다.

3.1. 목표 정의

기계학습 알고리즘 응용 결과 목표는 다음과 같다.

① 상품별 긍정·부정 비율
② 상품 중 가장 긍정·부정 적인 성향의 부품 종류
③ ②항목에서 선정된 부품의 세부 부품 종류
④ 상품 종류 및 세부 부품간의 관계도 분석

3.2. 데이터 수집 출처

기계학습 알고리즘 적용에 앞서 학습에 필요한 데이터 그룹의 수집 출처는 표 1과 같다.

표 1. 데이터 그룹 수집 출처
Table. 1 Data Collection Source(Reference)

Source		Entry
http://corearoadbike.com/	Category1 : 53,032	Title, Text
	Category2 : 24,798	
	Category3 : 27,935	
	Category4 : 41,706	
	Category5 : 23,630	
	Category6 : 23,740	
Total	194841	

수집된 데이터는 웹(Web)상에서 수집 가능한 거래 게시판 총 194,841개를 대상으로 데이터(구매, 판매)를 추출하였다.

3.3. 데이터의 이해

기계학습 알고리즘을 적용하는 단계 이전에 선행해야할 과정은 수집된 데이터를 일차적으로 가공해야 한다. 수집된 데이터는 최초에 각 알고리즘에 입력되는 변수들을 추출하는 과정이 실제로 전체과정(수집, 데이터준비, 알고리즘 적용, 결과추출, 결과분석)에서 60% 이상의 시간과 노력이 소모되었다. 또한 선별된 상품 거래 데이터에 대한 결과를 추출하기 위해서는 거래 상품 주제에 대한 높은 이해도가 요구되었다.

본 논문에서 사용된 거래 데이터는 로드바이크 중고 거래 데이터를 수집하였다. 로드바이크 상품에 관련된 결과를 추론하기 위해서 추출된 항목은 종류, 부품명과 종류, 부품 별 상관관계 등이다. 표 2는 로드바이크 준비 데이터 항목을 나타낸다.

표 2. 준비 데이터 항목
Table. 2 Data Entry(Prepared)

Entry	Description
Type	Competition, Endurance, Cycle Cross, Time Trial, Triathlon
Part Details	38 finished products Based on components
Relation	Wheel, Frame, Drive Train(Grade) Groups and set parts, etc.
Data	Season / off-season(In early December - end of March)

3.4. 기계학습 알고리즘 응용

본 논문에서 사용된 각 기계학습 알고리즘의 역할은 다음과 같다.

① Naïve Bayes : 긍정·부정 성향 판단을 위한 한글 단어, 로드바이크 관련 단어 빈도수와 통계치 추출(수치형)
② Decision Tree : Naïve Bayes에서 추출된 단어 별 직관적인 데이터 분류(명목형)
③ K-nearest neighbor : 분류된 데이터 그룹사이의 거리 추출(수치형)
④ Apriori : Decision Tree, Knn을 기반으로 핵심 키워드 연관 관계 분석(수치형)

3.4.1. Naïve Bayes Algorithm

수집 데이터에서 포함된 단어(제목, 본문의 긍정·부정 단어)의 빈도수와 통계치를 추출한다. 한글 단어 및 어휘는 형태소(국립국어원 표준국어대사전)에서 검색 결과로 나온 명사, 동사, 관형사·명사, 동사 등을 참고하여 긍정·부정 형태의 단어를 추출하였다. 다음은 단어의 긍정·부정을 판단하는 방법을 설명한다.

① 혼합 : 긍정·부정 단어가 혼합되어 있는 데이터인 경우 두 유형의 단어 통계치가 현재 데이터에서 60% 이상인 것을 선택(사회과학에서는 60%가 옳은 경우 성공적인 것으로 간주)
② 동일 : 긍정·부정 단어 비율이 동일한 경우 제목의 긍정·부정 통계가 비율이 높은 쪽을 판단하여 전체 통계에 가중치 10%(제목은 중요도가 높게 산정)
③ 공통 1 : 제목은 긍정·부정 통계가 비율이 높은 쪽을 판단하여 모든 과정에 통계에 가중치 10%
④ 공통 2 : 긍정·부정 유형의 데이터가 둘 다 존재하지 않는 경우 데이터는 단어추출 대상에서 제외한다. (사실상 없을 가능성이 매우 낮음)

3.4.2. Decision Tree Algorithm

나이브 베이스에서 추출된 단어는 수집된 데이터(194841건)에서 수많은 변수에 대한 경우의 수를 고려했을 때 고차원의 데이터를 생성한다. 의사결정 트리를 적용할 경우 고차원 데이터로 인해 과접합 가능성이 매우 높다.

본 논문에서는 3.1에서 정의한 목적을 기준으로 필요한 항목으로 데이터 분석을 진행한다. 이는 목적에 따라 알고리즘의 적용방법과 수집 데이터의 일차가공 등 수많은 변수가 존재(실제로 많은 변수들에 대한 예외처리 가능한 기계학습 알고리즘은 존재하지 않음)하기 때문에 범위를 제한하여 설정하였다.

긍정·부정 분류 이후 카테고리 1부터 6의 키워드 인식하여 최상위 부모로 정의한다. 각 카테고리의 데이터는 판매/구매로 분류되어 있으며, 추출된 세부 분류 항목은 긍정·부정 별 세부 상품으로 분류된다. 그림 1은 긍정·부정 부품 종류를 분류하는 과정을 나타낸다.

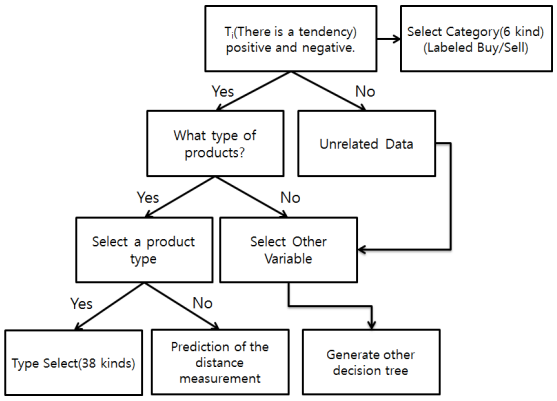

그림 1. 의사결정 트리 과정(긍정·부정)
Fig. 1 Decision Tree Process(Positive and Negative)

긍정·부정, 부품 종류가 추출되지 않은 데이터에 대해서는 준비 데이터에서 추가적으로 선택될 수 있는 다른 변수를 입력으로 새로운 의사결정 트리를 선택할 수 있다. 예로는 준비된 데이터 항목에 해당되지 않는 소모품명이나 가격, 무게 정보 등이다.

3.4.3. Knn(K-nearest neighbor) Algorithm

의사결정 트리 알고리즘으로 분류된 데이터 항목의 긍정, 부정, 세부 부품 단어의 발생 빈도수를 이용하여 거리를 측정한다. 유클리드 거리를 사용하여 계산하고, 추출된 값은 정규화 수행 이후, 거리가 가장 짧은 개수를 가지는 변수 정렬하고 연산을 반복 수행한다. 표 3은 긍정·부정, 부품 종류에서 나타난 브랜드를 기준으로 변수를 입력 예를 나타낸다.

위 예는 입력 변수의 종류와 결과 항목에 따라 추출되지 않은 물음표 항목(브랜드)에 대해 예측할 수 있다.

표 3. K-최근접 이웃 알고리즘 데이터 그룹
Table. 3 Knn Algorithms Data Group

Title	Positive	Negative	Part Type	Brand
T1	A	B	C	Brand 1
T2	A	B	C	Brand 2
T3	A	B	C	Brand 3
?	A	B	C	-

3.4.4. Apriori Algorithm

관계도출을 위한 연관분석은 앞에서 분류 알고리즘으로 추출된 분류 항목, 단어의 빈도수 및 확률 통계를 기반으로 Apriori와 응용한다. 연관 규칙에서 두 아이템 간의 관계의 강도를 측정하기 위해서 지지도와 신뢰도를 추출한다. 신뢰도는 추출된 부품 키워드 간에 연관 규칙에 대한 신뢰도를 나타낸다. 이는 거래데이터에서 특정 부품이 어느 부품과 관련성이 있는지 확인할 수 있다.

Ⅳ. 성능분석

실험환경은 데이터 처리/분석은 하둡(Hadoop)을 기반으로 파이썬(Python) 모듈을 이용하여 알고리즘을 수행한다. 각 기계학습 알고리즘은 연동되는 API나 프로그램이 없기 때문에 출력결과를 직접 계산하였다. 통계차트와 의사결정 트리의 데이터 시각화 부분은 D3 (Data-Driven Documents)를 이용하였다.

4.1. 입력 데이터 정의

다음은 성능평가를 위해 입력되는 데이터 예를 나타낸다.

① Naïve Bayes : 단어 추출(카테고리 1 - 6까지의 모든 단어 대상) : 긍정, 부정 : “명사”(기준 : 한국어 단어 60%이상 선택)

② Decision Tree : 세분화 타입 : 긍정·부정 별 부품 종류 선택, 기타 세부 부품, 관련 없음(기준 : 의사결정 트리에서 선택)

③ K-nearest neighbor : 비교 대상 : (목적에 따라 입력 변수 그룹이 다름), 정규화 범위 (0.00~10.00)

④ Apriori : 추출된 데이터 그룹 전체에서 특정 키워드 단어 존재 유무에 대한 지지도와 신뢰도

4.2. 출력결과

그림 2는 나이브 베이스에서 추출된 카테고리 별 긍정·부정 단어의 비율(방사형)을 나타낸다. 전체 카테고리에서 긍정·부정 성향 비율(평균 63%, 37%)은 긍정적인 것인 성향이 높았다. 카테고리 4의 경우 비율(49%, 51%)이 비슷하고, 카테고리 6의 경우 비율(63%)이 부정적인 성향이 높은 것으로 나타났다.

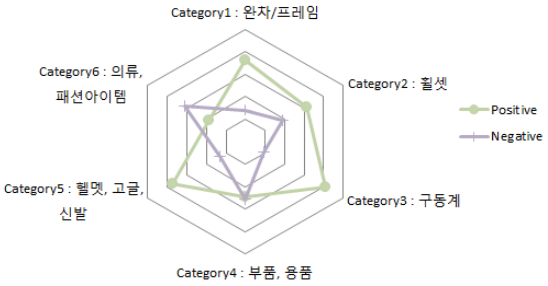

그림 2. 카테고리 별 긍정·부정 단어 비율 비교
Fig. 2 Positive, Negative Words Comparison Ratios by Category

그림 3, 4는 전체 카테고리에서 긍정·부정에 대한 비율을 나타낸 것이다.

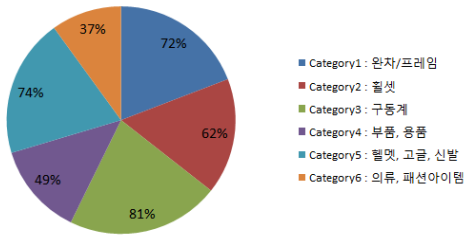

그림 3. 카테고리 별 긍정 단어 비율
Fig. 3 Positive Word Ratios by Category

분석 예) 긍정적인 성향이 가장 높은 비율로 나타나는 카테고리 3의 경우 판매자는 판매 전략상 어떤 구동계 종류에 대한 구매 만족도가 높은지 분석할 필요가 있다.

분석 예) 부정적인 성향이 높은 카테고리 6의 경우 사용자가 의류, 패션 아이템에 대한 제품의 만족도가 낮을 가능성이 크기 때문에 어떠한 상품 종류가 부정적인지 분석할 수 있다.

위 두 분석 요구사항은 공통적으로 카테고리를 세부 분류하고 내용을 확인해야 한다. 그림 5는 생성된 긍정·부정 별 카테고리 데이터 그룹을 의사결정 트리로 표현한 것이다.

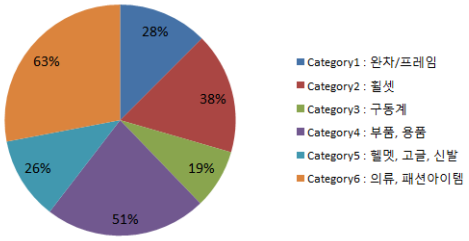

그림 4. 카테고리 별 부정 단어 비율
Fig. 4 Negative Word Ratios by Category

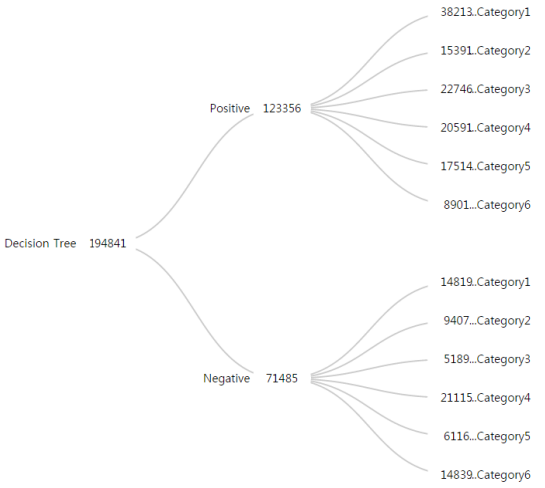

그림 5. 생성된 의사결정 트리(긍정·부정 및 카테고리 1-6)
Fig. 5 Generated Decision Tree(Positive and Negative, Category1-6)

카테고리 3을 예로 구매만족도 분석을 위한 항목은 구동계의 종류와 등급을 확인하는 것이다. 그림 6은 긍정 성향의 카테고리 3에서 세부 분류된 데이터 그룹을 나타낸다. 로드바이크 구동계는 일반적으로 3가지 종류로 분류되며 SHIMANO의 경우 7가지 등급으로 분

류된다.

생성된 의사결정 트리는 종류: SHIMANO, 등급: ULTEGRA의 분류 비율이 가장 높게 나타났다. 카테고리 3의 데이터 특징상 4가지 등급에 대한 키워드보다 부품에 대한 키워드가 데이터에 훨씬 많이 포함되었기 때문에 부품 키워드 인식을 선행한 후 이에 호환되는 등급과 부품 종류를 선택하도록 의사결정 트리를 작성하였다. 그림 6은 카테고리 3 데이터 그룹을 확장한 것이다.

카테고리 3에서 등급이 분류되지 않은 데이터는 전체 3459건, 세부 종류에서 1782건으로 이는 준비된 데이터 항목의 분류 선택 기준(정의된 변수의 범위)이 부족하여 나타나는 결과이다.

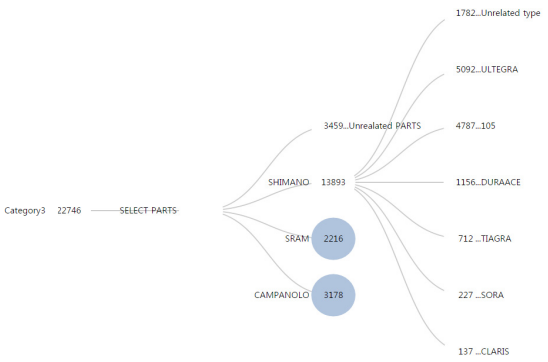

그림 6. 생성된 의사결정 트리(카테고리 3, 부품 종류와 등급)
Fig. 6 Generated Decision Tree(Category 3, Type and Grade)

모든 데이터 변수를 고려할 수 없기 때문에 분류되지 않은 항목에 대해서는 Knn을 이용하여 거리를 측정하고, 거리 값에 가까운 등급에 대해 예측한다. 표 4는 카테고리 3의 긍정·부정 데이터를 포함하는 등급 데이터의 거리(평균)을 나타낸다.

표 4. 데이터 거리(평균) – 부품 종류
Table. 4 Data Distance(Average) - Part Type

Data	Positive	Negative	Distance	Part Type
T1~Tn	1.3	2.0	7.3	SHIMANO
T1~Tn	2.2	0.9	1.8	SRAM
T1~Tn	3.1	1.8	2.34	CAMPANOLO
?	3.4	0.3	Prediction	Unrelated

데이터 간의 거리측정은 긍정·부정을 포함하는 전체 데이터에서 등급 항목에 대한 거리를 측정하였다. 측정되는 거리 값은 전체 데이터에서 측정한 데이터 거리에서 가장 가까운 거리를 찾는 것이 효율적이다. 분석 결과 관련되지 않은 데이터 유형은 가장 거리가 짧은 등급 : SRAM 으로 예상할 수 있다. 각 항목에 대한 키워드 관심도 측정을 위해 Apriori 을 이용하여 신뢰도를 측정하였다. 표 5는 키워드 SRAM과 등급 별 키워드에 대한 관심도를 분석하였다.

표 5. SRAM -> 등급 별 키워드(관심도)
Table. 5 SRAM -> Grade by Keyword(Interest)

Condition	Result	Approval rating	Confidence	Result
SRAM	RED22	0.01	0.5	4.613
SRAM	RED	0.01	0.5	15.139
SRAM	FORCE22	0.01	0.5	9.103
SRAM	FORCE	0.01	0.5	18.117
SRAM	RIVAL	0.01	0.5	6.847
SRAM	APECS	0.01	0.5	0.512

표 6은 카테고리중 긍정적 성향이 가장 높게 나타난 키워드 SHIMANO과 등급 별 키워드에 대한 관심도를 분석하였다.

표 6. SHIMANO -> 등급 별 키워드(관심도)
Table. 6 SHIMANO -> Rating by Keyword(Interest)

Condition	Result	Approval rating	Confidence	Result
SHIMANO	ULTEGRA	0.01	0.5	19.797
SHIMANO	105	0.01	0.5	29.247
SHIMANO	DURAACE	0.01	0.5	14.062
SHIMANO	TIAGRA	0.01	0.5	7.681
SHIMANO	SORA	0.01	0.5	3.180
SHIMANO	CLARIS	0.01	0.5	0.914

Condition 키워드에서 Result 키워드가 포함될 확률이 상대적으로 낮았기 때문에 지지도를 최소 0.01로 설정하고 최소 신뢰도를 0.5로 설정한 결과 SRAM의 경우 RED와 FORCE 등급에 대한 신뢰도가 가장 높게 나타났으며 SHIMANO의 경우 105, ULTEGRA 등급에

대한 신뢰도가 높은 순으로 나타났다. 앞의 예를 통합적으로 분석한 결론으로 카테고리 3에서 통계상 긍정성향이 높은 구동계 종류는 SHIMANO의 ULTEGRA 등급으로 나타났으며, 분류 데이터를 기준으로 관련이 없는 데이터를 예측했을 때 SHIMANO 다음으로 긍정적인 성향을 나타낸 구동계는 SRAM 구동계로 예측되었다. 이외 SRAM, SHIMANO 구동계 종류의 관심도 분석 결과 어떠한 등급에 사용자의 관심도(신뢰도)가 높다는 것을 확인할 수 있었다.

Ⅴ. 결 론

본 논문에서는 기계학습 알고리즘을 응용하여 사용자의 거래 성향을 분석 하였다. 성향 분석 결과 정의한 목적에 따라 다양한 결과를 추측할 수 있다는 것을 증명할 수 있었다.

그러나 수집된 데이터의 초기 데이터 가공에 많은 시간과 비용을 소비해야 했다. 이는 데이터의 주제와 목적에 따라 많은 변수가 존재하였기 때문이다. 실질적으로 기계학습을 위해 필요한 학습 데이터는 오랜 기간 동안 축적되어야 할 필요가 있고, 대부분 많은 변수들을 각 알고리즘 별 직접 입력해야하는 어려움이 존재했다.

기존 연구, 공개 및 상용 프로그램들은 데이터 수집에 대한 다양한 기능들을 제공하고 있다. 하지만 수집된 데이터를 입력데이터로 효율적으로 가공해주는 기능을 제공하는 데이터마이닝 연구 및 자동화 프로그램은 비교적 부족한 것으로 나타났다.

향후 연구로는 이러한 문제를 해결하기 위해 특정 기계학습 알고리즘에 적절한 수집된 데이터를 일괄적으로 정형화된 데이터로 가공할 수 있는 연구가 필요할 것으로 예상한다.

REFERENCES

[1] Jo-Hyeon, Park-Sangseon, “Understanding Product Satisfaction in the Context of Online Trading”, *Journal of the Korea Contents Association*, Vol. 13, No. 5, pp. 436-442, 2013.

[2] KCA, “Process Mining technology trends for Big Data analysis”, Korea Communication Agency, Information Communication Technology Issues & Prospect, 2014.

[3] Choi-Gyeyeong, “Artificial Intelligence: Disruptive innovation and evolution of the Internet platform”, Korea Information Society Development Institute, Premium Report, 2015.

[4] Greg Banks, “More growth options up front – Big data enables a new opening step in the growth decisionmaking process”, Deloitte Newsletter, 2014.

[5] Im-Sujong, Min-Okgi, “Machine Learning Technology Trends for Big Data Processing”, Electronics and Telecommunications Research Institute, Electronics and Telecommunications Trends, 2012.

[6] Lee-Byoungyup, Lim-Jongtae, Yoo-Jaesoo, “Utilization of Social Media Analysis using Big Data”, *The Journal of the Korea Contents Association*, Vol. 13, No. 2, pp. 211-219, 2013.

[7] McCallum, Andrew, and Kamal Nigam, “A comparison of event models for naive bayes text classification”, *AAAI-98 workshop on learning for text categorization*, 1998.

[8] Winkler, Robert L, “*Introduction to Bayesian inference and decision*”, Vol. 15, No. 4, pp. 938-939, 1973.

[9] BARROS, Rodrigo Coelho, et al, “A Survey of Evolutionary Algorithms for Decision-Tree Induction”, *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, Vol. 42, No. 3, pp. 291-312, 2012.

[10] Choi-Jonghu, et al. “Application of Data Mning Decision Tree”, Statistic Korea, *Journal of The Korean Official Statistics*, Vol. 4, No. 1, pp. 61-83, 1999.

[11] Altman, Naomi S. “An introduction to kernel and nearest-neighbor nonparametric regression”, The American Statistician, Vol. 46, No. 3, pp. 175-185, 1992.

[12] Hastie, Trevor, and Rolbert Tibshirani. “Discriminant adaptive nearest neighbor classification”, Pattern Analysis and Machine Intelligence, *IEEE Transactions on*, Vol. 18, No. 6, pp. 607-616, 1996.

[13] Agrawal, Rakesh, and Ramakrishnan Srikant. “Fast algorithms for mining association rules”, *Proc. 20th int. conf. very large data bases, VLDB*, Vol. 1215, pp. 487-499, 1994.

[14] Perego, Raffaele, Salvatore Orlando, and P. Palmerini. “Enhancing the apriori algorithm for frequent set counting”, *Data Warehousing and Knowledge Discovery*, Springer Berlin Heidelberg, pp. 71-82, 2001.

최도현(Do-hyeon Choi)

2008년 2월 : 동서울대학 컴퓨터소프트웨어 공학사
2010년 8월 : 숭실대학교 컴퓨터학과 석사
2010년 9월 ~ 현재 : 숭실대학교 컴퓨터학과 박사과정
※관심분야 : Mobile, Network Security, Virtualization, PKI

박중오(Jung-oh Park)

2000년 7월 : 성결대학교 컴퓨터공학과 졸업
2003년 3월 : 명지대학교 전자계산교육 석사
2011년 8월 : 숭실대학교 컴퓨터공학 박사
2013년 3월 ~ 현재 : 동양미래대학교 조교수
※관심분야 : PKI, Network Security, Cryptography