ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2013 June: 19(2): 39~54
http://dx.doi.org/10.13088/jiis.2013.19.2.039

다중모형조합기법을 이용한 상품추천시스템

이연정
동국대학교_서울 일반대학원 경영정보학과
(yeonjeong.lee@ntdtv.com)

김경재
동국대학교_서울 경영학부
(kjkim@dongguk.edu)

전자상거래의 폭발적 증가는 소비자에게 더 유리한 많은 구매 선택의 기회를 제공한다. 이러한 상황에서 자신의 구매의사결정에 대한 확신이 부족한 소비자들은 의사결정 절차를 간소화하고 효과적인 의사결정을 위해 추천을 받아들인다. 온라인 상점의 상품추천시스템은 일대일 마케팅의 대표적 실현수단으로써의 가치를 인정받고 있다. 그러나 사용자의 기호를 제대로 반영하지 못하는 추천시스템은 사용자의 실망과 시간낭비를 발생시킨다. 본 연구에서는 정확한 사용자의 기호 반영을 통한 추천기법의 정교화를 위해 데이터마이닝과 다중모형조합기법을 이용한 상품추천시스템 모형을 제안하고자 한다. 본 연구에서 제안하는 모형은 크게 두 개의 단계로 이루어져 있으며, 첫 번째 단계에서는 상품군 별 우량고객 선정 규칙을 도출하기 위해서 로지스틱 회귀분석 모형, 의사결정나무 모형, 인공신경망 모형을 구축한 후 다중모형조합기법인 Bagging과 Bumping의 개념을 이용하여 세 가지 모형의 결과를 조합한다. 두 번째 단계에서는 상품군 별 연관관계에 관한 규칙을 추출하기 위하여 장바구니분석을 활용한다. 상기의 두 단계를 통하여 상품군 별로 구매가능성이 높은 우량고객을 선정하여 그 고객에게 관심을 가질만한 같은 상품군 또는 다른 상품군 내의 다른 상품을 추천하게 된다. 제안하는 상품추천시스템은 실제 운영 중인 온라인 상점인 'I아트샵'의 데이터를 이용하여 프로토타입을 구축하였고 실제 소비자에 대한 적용가능성을 확인하였다. 제안하는 모형의 유용성을 검증하기 위하여 제안 상품추천시스템의 추천과 임의 추천을 통한 추천의 결과를 사용자에게 제시하고 제안된 추천에 대한 만족도를 조사한 후 대응표본 T검정을 수행하였으며, 그 결과 사용자의 만족도를 유의하게 향상시키는 것으로 나타났다.

논문접수일 : 2013년 06월 11일 논문수정일 : 2013년 06월 12일 게재확정일 : 2013년 06월 18일
투고유형 : 국문일반 교신저자 : 김경재

1. 서론

전자상거래의 폭발적 증가는 소비자에게 많은 구매 선택의 기회를 제공한다. 이러한 상황에서 의사결정에 확신이 부족한 소비자들은 의사결정 절차를 간소화하기 위해 추천을 받아들인다. 직접적으로 물품을 살펴보거나 판매자와 상호작용이 어려운 전자상거래에서 추천의 역할은 더욱 커진다. 따라서 온라인 상점 추천시스템은 방대한 정보로 인한 온라인 쇼핑이 부담스런 고객에게 쇼핑 가이드라인을 제시하고 구매결정을 돕는 전문가로서의 역할을 할 수 있다. 즉, 추천시스템은 고객의 취향이나 선호를 바탕으로 고객에게 가장 가치 있는 제품이나 서비스를 찾도록 도와주는 가이드 역할을 하는 시스템이다. 따라서 정교한 추천시스템을 갖춘 온라인 상점은 다른 기업에 비해 고객관계관리에서 경쟁력을 가질

* 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업지원을 받아 수행된 것임 (No. 2010-0025689).

수 있다. 실제로 많은 온라인 상점들이 추천시스템을 도입하여 사용하고 있고, 대표적 예로는 amazon.com, barnesandnoble.com, netflix.com, cdnow.com 등이 있다.

추천시스템의 기업측면 가치를 살펴보면 상점을 둘러보고 있는 고객에게 적절한 상품을 추천함으로써 실구매자로 유도하거나, 고객이 깨닫기 전에 필요한 물품을 알려줌으로써 교차판매 기회를 잡을 수 있다. 이러한 서비스를 받게 된 고객은 편의성과 친밀감을 느끼게 되고, 이에 따라 고객의 충성도는 한층 높아질 것이다. 고객측면의 가치는 관심 제품이나 서비스를 찾기 위한 탐색 비용을 줄여주고 구매의 효율성을 제고한다. 이에 따라 고객의 취향에 맞는 보다 정확한 상품추천을 할 수 있는 추천알고리즘 개발의 중요성이 점점 높아지고 있다.

본 논문에서는 전자상거래의 폭발적 성장에 따른 일대일 마케팅 기법 실현도구로써 개선된 개인화 상품 추천시스템을 제시하고자 한다. 먼저, 본 연구에서는 문헌연구를 통해 일반적인 상품추천시스템의 문제점을 파악하여 이를 보완할 수 있는 개선된 상품추천시스템을 제시하고자 한다. 선행 연구에서 제안된 상품추천시스템 구축들은 다양한데, 대부분 협업필터링(collaborative filtering) 기법(Resnick et al., 1994)을 많이 활용하고 있으며, 이는 내용기반(content-based) 기법과 함께 추천시스템 기법의 두 축을 이룬다. 그러나 이 두 가지 방법은 고객과 고객 사이 또는 상품과 상품 사이의 연관성을 기반으로 추천을 하는 방식으로 구매가 상대적으로 적은 고객이나 구매가 상대적으로 적게 이루어지는 상품에 대해서는 추천이 불가능하거나 정교한 추천을 하지 못할 수 있으며, 이는 추천시스템 연구에서 희박성(sparsity)의 문제라고 하는 중요한 문제점이다. 또한 협업필터링의 경우에는 거래 데이터가 증가함에 따라 추천에 필요한 유사고객을 찾기 위한 유사성 계산에 필요한 연산량이 기하급수적으로 증가하는 문제가 발생하며, 이는 추천시스템 연구에서 확장성(scalability)의 문제라고 하는 매우 중요한 문제점 중에 하나이다. 따라서 협업필터링이나 내용기반 추천기법은 거래가 빈번하게 일어나지 않는 중소 또는 사업초기의 인터넷 쇼핑몰의 경우에는 추천의 성능이 좋지 못할 가능성이 있다. 반면, 거래가 빈번하게 발생하는 대형 인터넷 쇼핑몰에서는 추천의 효율성 면에서 문제가 발생할 수 있다(Kim and Kim, 2005). 또한, 협업필터링의 경우에는 반드시 상품에 대한 고객의 평가가 사전에 이루어져야 한다는 단점도 있다.

본 연구에서는 상기의 기존 상품추천시스템의 한계점을 보완할 수 있는 데이터마이닝 기반의 추천기법을 제안한다. 본 연구에서 제안하는 기법은 고객의 프로필이나 상품거래자료와 같이 이미 구축되어 있는 데이터를 활용하여 데이터마이닝 기법을 이용하여 추천을 하므로 상품에 대한 고객의 사전 평가자료를 필요로 하지 않는 장점이 있다. 또한, 한 번의 추론을 통해 추천을 위한 규칙을 도출하면 매 추천 때마다 유사성 연산을 할 필요가 없으므로 협업필터링 기법 등에서 발생하는 확장성, 희박성 문제를 어느 정도 완화할 수 있다. 데이터마이닝을 이용한 추천시스템 개발에 관한 연구도 이미 이루어졌으나, 대부분 선행연구에서는 단일 데이터마이닝 기법을 이용한 추천기법에 관한 연구들이 대부분이었다. 본 연구에서는 여러 개의 추천기법의 결과를 융합하여 사용하는 다중모형조합기법을 이용하는 추천기법이므로 단일 방법을 이용하는 기존 연구보다 추천 성능을 개선할 가능성이 있다. 즉, 본 연구에서 제시하는 모형은 다중모형조합기법을 통해 데이터마이닝 분류기법의 과잉학습 단점을 보완하고 실제 값과 예측 결과 값의 차이를 줄여 모형 정확도를 높일 것으로 기대된다. 다중모형조합기법

은 여러 개의 분류모형을 조합하여 각 모형의 단점을 최소화하면서 분류정확도를 제고하기 위해 많이 사용된다. 다중모형조합기법을 통해 단일 모형이 가지는 단점이나 편차를 다른 모형의 결과를 이용하여 보완할 수 있다는 장점이 있기에 많은 연구자들에 의해서 데이터마이닝 문제에 적용되어 왔다. 본 연구에서는 실제 온라인 상점 구매데이터를 적용하고 사용자의 만족도를 확인하여 제안하는 모형의 실제 적용가능성을 검증할 것이다.

본 연구는 다음과 같이 구성되어 있다. 제 2장에서는 추천시스템에 대한 선행연구와 본 연구에서 사용되는 다중모형조합기법에 대한 간략한 설명이 제시된다. 제 3장에서는 본 연구에서 제안하는 추천시스템에 대해 설명하고, 제 4장에서는 본 연구에서 제안하는 시스템의 유용성을 검증하기 위한 데이터와 실험설계에 대해 설명한다. 제 5장에서는 추천시스템 성능에 대한 유용성 검증결과를 제시하고, 마지막 제 6장에서는 연구의 결론과 한계점을 제시한다.

2. 선행연구

본 연구에서는 전술한 바와 같이 데이터마이닝과 다중모형조합기법을 통해 기존 상품추천시스템의 문제점을 보완한 모형을 제시하고 한다. 본 장에서는 먼저 상품추천시스템에 대한 선행연구들을 살펴보고, 본 연구에서 제안하는 상품추천시스템의 기반 기술인 데이터마이닝 기법을 살펴 볼 것이다. 본 연구에서는 인공신경망, 의사결정나무, 로지스틱 회귀분석과 같은 전통적인 데이터마이닝 기법이 사용되지만 이에 대해서는 이미 많은 선행연구에서 소개하고 있으므로 설명을 생략하고 다중모형조합기법과 연관규칙기법에 대해서만 간략히 살펴본다.

2.1 전통적인 추천기법과 한계점

상품추천시스템에서 가장 중요한 부분은 추천기법이라 할 수 있으며, 내용기반 필터링과 협업필터링이 가장 대표적인 기법이다. 내용기반 필터링은 상품 간의 속성 관계를 이용하여 고객이 선호했던 상품과 유사한 속성을 가진 상품을 추천하는 방식이다. 이 기법의 가장 중요한 장점은 상품 자체의 속성을 이용하여 추천하는 방식이기에 직접적이고 단순하다는 점이다(Wu et al., 2001). 그러나 이 기법은 추천의 대상이 되는 상품의 속성을 미리 추출해 두어야 이용 가능한데, 신상품이 지속적으로 출시되는 상황에서는 속성 추출이 용이하지 않다는 단점이 있다. 또한, 이 기법은 고객이 이전에 평가하거나 구매한 상품과 관련된 상품을 추천하는 방식이기에 고객의 과거 행태에 지나치게 의존한다는 한계점도 있다. 이러한 한계점들로 인해 최근의 추천시스템 연구에서는 협업필터링이 더 활발하게 연구되고 있다(Ahn et al., 2006).

협업필터링은 고객 간의 유사성을 바탕으로 추천하는 방식이다. 즉, 추천을 하려는 고객과 유사한 구매행태를 보이는 다른 고객의 속성을 토대로 추천하는 방식이다. 협업필터링에 대한 초기연구로는 Tapestry(Goldberg et al., 1992), GroupLens(Resnick et al., 1994)의 사례가 있으며, Ringo와 Video Recommender 등의 이메일과 웹 기반 협업필터링에 의한 상품추천시스템이 개발된 바 있다(Sarwar et al., 2000). 협업필터링은 일반적으로 평가자들이 동질적인 평가를 하는 상품군에 대하여 상대적으로 높은 추천성과를 보이는 것으로 알려져 있으며, 평가 데이터가 충분한 상품의 경우에도 다른 추천기법에 비해 상대적으로 좋은 추천성과를 보이는 것으로 알려져 있다(Konstan et al., 1997; Pazzani, 1999). 그러나, 기본적으로 협업필터링은 상품에 대

한 고객 평가자료를 기반으로 추천을 하는 방식이므로 이를 많이 보유하고 있는 대형 인터넷 쇼핑몰에서는 유용하지만, 평가자료의 양이 비교적 적은 중소 인터넷 쇼핑몰 또는 사업 초기단계에 있는 인터넷 쇼핑몰의 경우에는 성과가 좋지 못할 가능성이 있다(Ahn et al., 2006). 이러한 한계점은 여러 선행연구에서 협업필터링이 가진 가장 심각한 문제점 중 하나로 지적되고 있으며, 이를 추천시스템 연구에서는 흔히 희박성(sparsity) 문제라고 한다(Kim et al., 2002a; Cho et al., 2002; Kim et al., 2002b; Kim et al., 2003; Kim et al., 2004; Cho et al., 2004; Cho and Kim, 2004; Kim et al., 2005; Kim and Yum, 2005; Adomavicius and Tuzhilin, 2005; Kim and Kim, 2005; Ahn et al., 2006, Kim and Ahn, 2011 참고). 선행연구에서는 이 한계점을 극복하기 위해 웹 로그 자료를 활용하여 부족한 평가자료를 보완하고자 하는 노력을 하였다(Cho et al., 2002; Kim et al., 2002a; Kim et al., 2003; Cho and Kim, 2004; Kim et al., 2005). 그러나 일반적으로 웹 로그 자료는 대용량이고 정제되지 않은 형태로 이루어져 있기에 전처리를 하는 데에 많은 시간과 비용이 소요된다는 단점이 있다. 다른 연구에서는 사용자가 평가하지 않은 상품이나 신규 고객에 대해서 구매거래 데이터에 사회연결망분석의 개념을 적용하여 도출된 중심상품 또는 중심고객과의 관련성을 활용하여 구매가능성을 예측한 후 추천하는 방법을 제안하였다(Park et al., 2009, Cho and Bang, 2009). 그러나 이 방법의 경우에도 거래 데이터를 추가적으로 확보하여야 하는 어려움이 있고, 구매가능성을 예측해야 하는 새로운 예측문제가 발생하므로 어려움이 있다.

협업필터링의 다른 중요한 한계점 중 하나는 고객의 수와 구매거래 데이터가 증가함에 따라 유사한 고객을 찾기 위한 필터링 연산량이 기하급수적으로 증가하여 시간과 비용이 과다하게 발생할 수 있다는 것이며, 선행연구에서는 이러한 한계점을 확장성(scalability) 문제라고 한다(Kim et al., 2002a; Cho et al., 2002; Kim et al., 2002b; Kim et al., 2003; Cho et al., 2004; Cho and Kim, 2004; Kim et al., 2005; Adomavicius and Tuzhilin, 2005, Kim and Kim, 2005; Ahn et al., 2006, Kim and Ahn, 2011 참고). 이 한계점은 신속한 대응을 요구하는 인터넷 사용자의 특성을 감안할 때 고객의 이탈을 이끌 수 있는 매우 심각한 단점이다. 선행연구에서는 이 문제점을 보완하기 위해서 여러 방법을 제안한 바 있다.

대표적인 예로는 Kim et al.(2002a), Cho et al.(2002), Kim et al.(2002b), Cho et al.(2004), Cho and Kim(2004), Kim et al.(2005) 등이 제안한 방법으로, 상품계층도(product taxonomy)를 활용하는 방법이 있다. 그러나, 이 방법은 상품계층도 내 각 상품계층군 안에서는 고객의 선호도가 제대로 반영되지 않기에 추천의 성과가 떨어질 가능성이 있다. 또한, 상품계층도의 작성이 추천 성과에 큰 영향을 미칠 수 있는데 선행연구에서는 전문가의 주관적인 판단을 이용하는 방식을 제안하였으나 이 점 역시 단점이 될 수 있다.

확장성 문제를 완화하기 위한 방법으로는 군집분석을 협업필터링의 사전과정으로 수행하여 탐색공간을 줄이는 방법이 있다. Kim et al.(2003)은 K-평균 군집분석을 협업필터링 사전과정으로 이용하여 탐색공간을 줄이고자 하였고, Roh et al.(2003)과 Kang(2003)은 군집분석의 일종인 자기조직화지도를 이용하여 탐색공간을 축소하였다. 그러나 이 같은 선행연구들은 추천을 위해 협업필터링 외에 추가적인 분석을 수행하여야 하는 어려움이 있다.

이상의 선행연구 결과들을 종합해 보면 협업필터링이 상품추천에 있어서 유용한 방법이지만, 희박성과 확장성의 문제가 한계점이며, 이러한 한계점이

보완되지 않으면 추천의 성과가 저하될 수 있다는 점을 공통적으로 지적하고 있다. 본 연구에서는 상기의 기존 상품추천시스템의 한계점을 보완할 수 있는 데이터마이닝 기반의 추천기법을 제안한다. 본 연구에서 제안하는 모형은 한번의 추론을 통해 추천을 위한 규칙을 도출하면 매 추천 때마다 유사성 연산을 할 필요가 없으므로 협업필터링 기법 등에서 발생하는 확장성, 희박성 문제를 어느 정도 완화할 수 있다. 또한, 고객의 프로필이나 상품거래자료와 같이 이미 구축되어 있는 데이터를 활용하여 데이터마이닝 기법을 이용하여 추천을 하므로 상품에 대한 고객의 사전평가자료를 필요로 하지 않는 장점이 있다. 본 연구에서는 여러 개의 데이터마이닝 기법의 결과를 융합하여 사용하는 다중모형조합기법을 이용하는 추천기법이므로 단일 방법을 이용하는 기존 연구보다 추천성능을 개선할 가능성이 있다.

2.2 다중모형조합기법(Ensemble Method)

다중모형조합기법은 모형의 성능을 향상시키기 위해 다수개의 모형결과를 조합하여 최종적인 결론을 내리는 기법이다. 다중모형조합기법은 인공신경망 등 인공지능기계학습 분류모형에서 모형이 주어진 학습데이터에 대한 적합 능력이 과하게 뛰어나 발생하는 '과잉학습' 문제의 해결방안이 될 수 있다. 또한 모형의 예측 결과 값과 실제 데이터의 결과값의 오차를 줄여 분류모형 정확도를 향상시킨다. 다중모형조합기법은 여러 가지 방법이 있는데 본 연구에서는 대표적인 기법인 Bagging과 Bumping을 활용한다.

Bagging은 Leo Breiman에 의해 제안되었다(Breiman, 1994, 1996). Bagging은 "Bootstrap aggregation"의 약자이며, 기계학습기법에서 분류나 예측의 성과와 안정성 제고를 위하여 여러 개의 기계학습기법의 결과를 조합하는 방법이다. 이 방법은 편차를 감소시키고 과대적합을 피하게 하는 역할도 한다. 일반적으로 이 방법은 의사결정나무 분석에 사용되지만 어떤 형태의 모델에도 적용 가능하며, 일반적으로 알려진 모형결과 평균화기법의 특별한 형태라고 할 수 있다. 이 방법은 여러 분류기의 결과를 평균화하는 역할을 하기 때문에 선형 모형의 성과를 개선하는 데에는 적합하지 않을 수 있다. 또한 이 방법은 k최근접이웃기법과 같이 매우 안정된 모형의 성과를 개선시키는 것도 어려운 것으로 알려져 있다(http://en.wikipedia.org/wiki/Bootstrap_aggregating).

Bumping은 "Bootstrap umbrella of model parameters"를 의미하는 것으로 Bagging이 개별 분류기의 결과를 고려하지 않고 모든 분류기의 평균화된 값을 이용하는 것에 반해, Bumping은 모든 분류기의 결과를 고려하는 것이 아니라 가장 낮은 오차의 결과값을 갖는 모형의 결과값만을 사용한다(Tibshirani and Knight, 1995; Heskes, 1997; Lee and Kwak, 1999).

2.3 연관규칙기법

연관규칙기법은 두 객체의 발생빈도가 서로 연관되어 있다고 판단될 때 그 연관정도를 파악할 수 있는 분석기법이다. 특히, 연관규칙기법에서 가장 폭넓게 사용하고 있는 장바구니분석은 상품 구매 시 일련의 제품이 있다고 할 때 같이 구매되는 다른 일련의 제품이 함께 존재할 것인지를 측정하는 체계적 기법이다. 이 기법은 상품 간의 연관성 분석을 통해 상품의 동시판매가능성 등을 측정할 때 이용한다. 장바구니분석에서는 지지도, 신뢰도, 향상도의 지표를 이용하여 연관관계 규칙을 분석한다.

지지도(Support)는 전체거래 중 상품 A와 상품 B를 동시에 포함하는 거래의 비율을 의미하며, 연

관규칙의 중요도를 판단할 때 이용된다. 신뢰도(Confidence)는 상품 A를 포함하는 거래 중에서 상품 B가 포함된 거래의 비율을 의미하며, 두 상품 간의 연관성의 강도를 측정하는 데에 이용된다. 마지막으로, 향상도(Lift)는 상품 A를 구매한 경우, 그 거래가 상품 B를 포함하는 경우와 상품 B가 임의로 구매되는 경우의 비율을 말하는 것이며, 이 지표는 두 상품의 연관성이 우연적인 것인지를 판단할 때 이용된다.

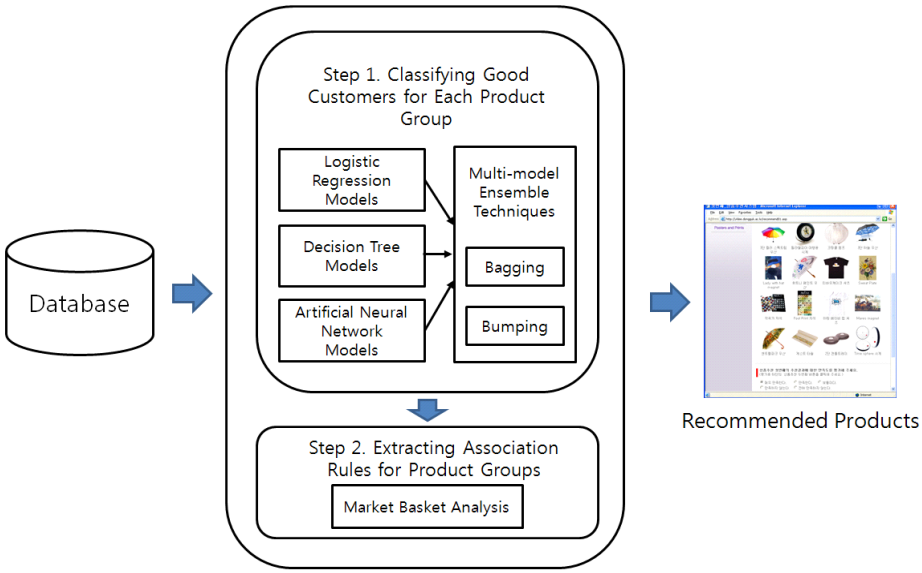
3. 제안 상품추천시스템

기존연구를 검토해 본 결과 많은 연구자들이 추천 알고리즘을 제안하고 있지만 대부분의 연구가 협업 필터링이나 의사결정나무와 같은 단일 기법을 활용하고 있다. 최근 데이터마이닝 연구들을 살펴 보면 단일기법보다는 다양한 기법의 조합모형이 더 우수한 성과를 보이는 것으로 보고되고 있다. 다양한 기법의 조합모형이 가진 장점이 있기에 이를 바탕으로 본 연구에서는 기존의 추천 알고리즘을 보완하여 보다 개선된 추천 알고리즘을 제안하고자 한다.

본 연구에서 제안하는 상품추천시스템에서는 기존 연구들처럼 온라인 상점 고객에 대해 구매기록이 있는 고객과 구매 기록이 없는 고객으로 분류하여 분석하는 것이 아니라 구매기록이 있는 고객 중 특정 상품군의 상품을 구매한 고객에 대한 분석을 한다.

본 연구에서 제안하는 상품추천시스템은 데이터마이닝에서 분류모형에 자주 활용되는 로지스틱 회귀분석, 인공신경망, 의사결정나무를 사용하여 분류모형을 개발하고 이의 결과를 다중모형조합기법을 활용하여 결합하여 우량고객을 선발한 후, 이들의 구매행동을 연관규칙을 활용하여 분석하고, 우량고객이 동시에 자주 구매하는 상품을 추천하는 방식이다. 본 연구에서 제안하는 연구모형은 <Figure 1>과 같다.

본 연구에서는 세 개의 분류모형들의 결과 값을



<Figure 1> Proposed Model of Product Recommender System

다중모형조합기법의 'Bumping'과 'Bagging'으로 조합하여 이 중 오차율이 낮은 모형을 채택한다. 이는 분류모형의 과잉학습을 완화시켜 보다 견고한 분류모형 결과를 생성하게 하며 데이터의 실제값과 모형의 예측 결과값의 오차범위를 축소할 수 있을 것으로 기대된다.

기존 연구에서는 상품추천 단위로 개별상품을 활용하여 추천결과를 생성하였으니, 본 연구에서는 상품군 단위로 추천결과를 생성한다. 이는 기존 추천시스템의 '판매자 의도 반영 불가능'이란 한계점을 다소 해소할 수 있는 기회제공을 하며 실제 온라인 상점의 빈번한 상품구성 변동에 탄력적인 추천결과를 생성하게 한다. 또한 내용기반 추천방식에서 단점으로 지적된 '다양한 추천의 어려움'이란 단점을 완화시킬 수 있다.

본 연구에서 제안하는 추천시스템을 구현하기 위해 다음과 같은 작업을 하였다. 고객 데이터에 대한 결측치 처리, 필드명 변경 등의 전처리 과정을 거쳐 각 모형에 적합한 변수를 선정하게 된다. 로지스틱 회귀분석의 입력변수로 사용할 연속형 변수의 경우 독립표본 T 검정을 활용하고, 범주형 변수의 경우 카이제곱 검정을 실시하여 5% 유의 수준에 유의한 변수들만 선정하여 모형을 구축한다.

4. 실험

4.1 실험 데이터

본 연구에서 제안하는 상품추천시스템의 유용성을 확인하기 위하여 실제 데이터를 활용한 간단한 시스템 프로토타입을 구축하고 이를 활용하여 실제와 유사한 환경에서의 추천 효과를 확인한다. 이 실험에서는 실제 인터넷에서 운영되고 있는 약 550종의 유명 미술관과 박물관의 상품을 취급하는 온라인 상점 "I아트샵"으로 부터 수집된 구매데이터를 사용한다. 수집된 구매 데이터는 총 5759 개의 데이터이나 연구자가 결측치 처리를 하여 8개 필드의 3169개 레코드를 연구데이터로 사용하게 되었다.

<Table 1> Selected Input Variables and Description after Preprocessing

Variables	Description	Value
id	Customer ID	Text
age	Customer Age	Integer
gender	Customer Gender	1: Female / 2: Male
job1~job13	Artists, Office Jobs, Educational Jobs, House Wives, Unemployed, Administrative Jobs, Marketers, Others, Special Jobs, Sales, Service Jobs, Students, Technicians	0: OK / 1: N/A
zip1~zip38	Dobong-Gu, Dongdaemun-Gu, Dongjak-Gu, Youngdeungpo-Gu, Chung-Gu, Chungrang-Gu, Jongro-Gu, Kangbuk-Gu, Kangdong-Gu, Kangnam-Gu, Kangseo-Su, Guro-Gu, Kwangjin-Gu, Kwanak-Gu, Mapo-Gu, Rowon-Gu, Sungbuk-Gu, Seocho-Gu, Seodaemun-Gu, Sungdong-Gu, Songpa-Gu, Yangchun-Gu, Yongsan-Gu, Eunpyung-Gu, Kangwon-Do, Kyunggi-Do, Keungsangnam-Do, Keungsangbuk-Do, Jeju-Do, Jeunranam-Do, Jeunrabuk-Do, Chungcheungnam-Do, Chungcheungbuk-Do, Kwangju-Si, Daegu-Si, Daejun-Si, Busan-Si, Ulsan-Si, Incheun-Si	
F	Customers of Fashion Product Group	
HD	Customers of Home Decoration Product Group	
O	Customers of Office Supplies Product Group	
P	Customers of Poster Product Group	

‘조각상’ 상품군의 경우 레코드의 수가 2개로 실험을 하기에 그 수가 현저히 낮아 제외시켰다. 범주형 변수를 통계분석에 적합하도록 더미변수(dummy variable)화하고 이상치를 제거하는 등 전처리 작업을 끝낸 결과 필드 8개, 레코드 3167개의 데이터 셋으로 정리하였다. <Table 1>은 전처리 후 변수와 내역이다.

I아트샵의 현재 회원수는 약 10만 명이며, I아트샵의 상품은 그 특성에 따라 조각상, 오피스, 홈데코, 포스터의 5개 상품군으로 분류하였다. 조각상 상품군은 거래수가 2건으로 연구 데이터에서 제외하였으므로 이는 고려하지 않는다. <Table 2>는 ‘I아트샵’의 상품 분류체계이다.

<Table 2>에서 1단계 Fashion 상품군은 총 5개의 2단계 상품군으로 구성되어 있다. 2단계 상품군은 열쇠고리, 타이와 스카프종류, 셔츠류, 우산, 가방류이다. 1단계 HomeDeco 상품군은 4개의 2단계 상품군으로 구성되어있으며 HomeDeco의 2단계 상품은 목욕용품, 인테리어 소품, 디자인자석류, 각종 테이블웨어이다. Office 2단계 상품군은 달력, 카드, 메모지, 취미용품, 편지지, 노트 등이다. Poster 상품군의 2단계 상품군은 단일 구성이다.

<Table 2> Product Taxonomy for ‘I Artshop’

Tier 1 Product Group	Codes for Tier 2 Product Group	Products
Fashion	F_keyholder	Key Holders
	F_ties	Ties
	F_tshirt	T-shirts
	F_umbrella	Umbrellas
	F_bag	Bags
Home Deco.	HD_bath	Bathroom Wares
	HD_interior	Interior Properties
	HD_magnet	Magnets
	HD_tblware	Table Wares
Office	O_calendar	Calendars
	O_card	Cards
	O_cube	Memo Pads
	O_diary	Diaries
	O_hobby	Hobby Items
	O_letter	Letterhead Stationaries
	O_note	Notes
	O_etc	Other Items
Poster	P_picture	Posters

실험 결과를 토대로 상품 추천시스템 프로토타입을 구축한다. 프로토타입의 유용성은 온라인 설문을 통해 고객의 만족도를 확인하는 방식으로 확인한다. 보다 상세한 실험설계에 대해서는 다음 절에서 설명한다.

4.2 실험설계

본 연구에서 제안하는 추천 알고리즘은 다중모형 조합기법을 이용하며, 이를 구성하는 단일기법은 로지스틱 회귀분석, 의사결정나무, 인공신경망 모형이다. 이 중 의사결정나무는 C4.5를 사용하였고, 인공신경망 모형은 오류역전파 전방향 인공신경망을 사용하였다.

모형 구축에 사용되는 데이터의 구성에 있어서 로지스틱 회귀분석과 의사결정나무에서는 학습용 데이터와 검증용데이터를 8:2의 비율로 구성하고, 인공신경망은 학습용, 검증용, 테스트용으로 6:2:2의 비율로 실험한다. 데이터 분류는 임의추출방식에 의한다. <Table 3>은 모형별 데이터의 수를 정

<Table 3> Data Sets for Models

Classification Model	Total	Training Set	Validation Set	Test Set
Logistic Regression	3167	2533	634	N/A
Decision Trees	3167	2533	634	N/A
Artificial Neural Networks	3167	1901	634	632

리한 것이고 <Table 4>는 상품군별 데이터 분포를 나타낸 것이다.

상품군 선호고객에 대한 규칙은 세 개의 분류모형인 로지스틱 회귀분석 모형, 의사결정나무 모형, 인공신경망 모형의 검증용 데이터 결과값을 다중모형조합기법으로 각각 결합한 결과 중 우수한 모형의 결과를 토대로 생성된다. 선호고객에게 추천할 상품을 제시하는데 이용할 상품 간 연관규칙은 장바구니분석기법을 이용하여 생성한다. 실험은 SAS E-Miner를 이용하였으며, 본 연구의 목표는 다중모형조합기법의 유용성을 확인하기 위한 것이므로, 단일 모형에 대해서는 특별한 실험 설정을 하지 않고 실험 패키지에서 기본적으로 제공하는 설정을 이용하였다.

<Table 4> Data Sets for Product Sets

	Total	Training Set	Validation Set
Fashion	1176	941	235
Home Deco.	405	324	81
Office	1368	1094	274
Poster	218	174	44

5. 실험결과

실험설계에 제시된 절차에 따라 실험 한 결과 중 상품군별 선호고객 분류모형의 결과는 <Table 5>와 같다.

<Table 5>에서 제시된 것과 같이 세 개의 분류모형의 결과, 패션상품군에서는 로지스틱 회귀분석이 가장 좋은 예측결과를 나타냈고, 홈데코와 오피스상품군에서는 의사결정나무가 가장 높은 예측률을 보였으며, 포스터상품군에서는 인공신경망이 가장 낮은 오차를 나타냈다. 세 분류모형의 Bagging과 Bumping으로 조합한 결과는 포스터상품군을 제외한 3개 상품군에서 Bumping의 결과가 우수하게 나타났다. 다중모형조합 결과는 단일모형의 상품군별 우수 모형에 비해 고르게 높은 예측률을 보였으나 포스터상품군은 예외였다. 상품 간 연관규칙 결과는 장바구니 분석기법의 규칙의 대표성을 나타내는 지지도(Support)가 높으며 향상도(lift)가 1점 이상인 규칙만을 채택하여 규칙 31개를 도출하였다. 이 중 중복되는 규칙을 제거한 후 15개의 규칙이 최종적으로 생성되었다. 15개 규칙 중 11개의 규칙이 오피스상품군의 상품 간의 연관관계를 이루고 있으며 나머지 4개의 규칙 또한 오피스상품군, 홈데코상품군, 오피스상품군과 패션상품군의 연관관계 규칙이었다. 도출된 연관규칙에 대한 상세한 내용은 <Table 6>에 정리하였다.

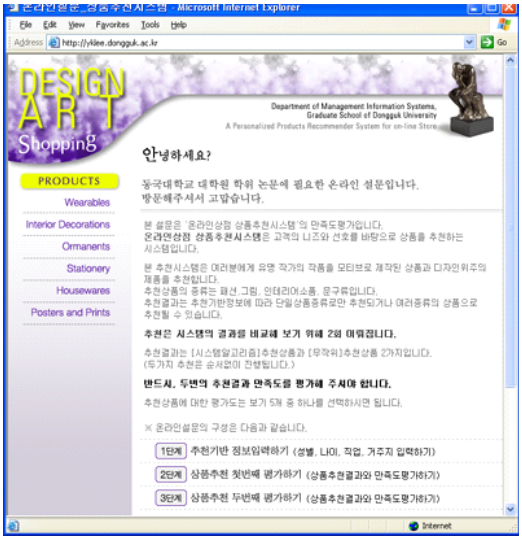
<Table 5> Classification Results of the Preferred Customers for Each Product Set

	Fashion	Home Deco.	Office	Poster
Logistic Regression	64.67%	87.38%	56.94%	93.06%
Decision Trees	63.72%	87.70%	58.04%	93.06%
Artificial Neural Networks	63.09%	87.22%	57.89%	93.20%
Bagging	63.81%	87.40%	57.63%	93.10%
Bumping	65.14%	88.16%	58.13%	93.06%

본 연구의 추천 결과의 유용성을 확인하기 위해 추천 결과를 실제 이용자들에게 제공하고 이에 대한 이용자들의 만족도를 확인하기 위해 시스템 프로토타입을 구축하였다. 프로토타입은 ASP와 Java Script로 구현되었으며, 데이터베이스 시스템은 Microsoft Access를 기반으로 한다. 구축된 프로토타입의 운영과정은 2단계로 구성된다. 1단계는 추천을 위해 기본적으로 필요한 고객의 인적사항 등 기반정보를 입력 받고 2단계는 제안추천시스템 모형의 추천결과에 대한 평가와 무작위 추천결과에 대한 평가를 받는다.

<Table 6> Extracted Association Rules

	Lift	Support(%)	confidence(%)	Extracted Association Rules
1	2.77	1.84	19.35	O_cube ==> HD_magnet
2	2.58	2.07	18	O_etc ==> HD_magnet
3	2.54	1.84	16	O_etc ==> O_hobby
4	2.51	3.6	23.86	O_note ==> O_cube
5	1.94	1.61	14	O_etc ==> O_card
6	1.9	3.3	28.67	O_etc ==> O_note
7	1.75	2.68	16.67	O_diary ==> O_cube
8	1.7	1.61	10.66	O_note ==> O_hobby
9	1.69	1.84	12.18	O_note ==> O_card
10	1.67	4.06	26.9	O_note ==> O_diary
11	1.57	2.91	25.33	O_etc ==> O-diary
12	1.54	1.69	14.67	O_etc ==> O_cube
13	1.5	1.69	10.48	O_diary ==> HD_magnet
14	1.45	1.69	10.48	O_diary ==> O_card
15	1.17	1.76	28.05	O_hobby ==> F_bag

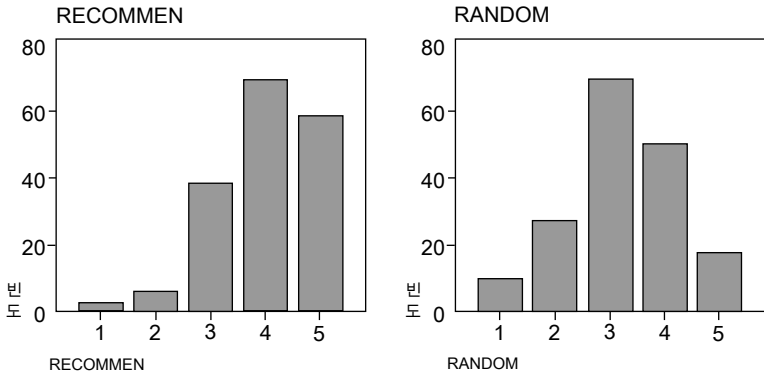


<Figure 2> Screenshot of the System Prototype

<Figure 3> Example of Product Recommendation

<Figure 2>는 구축된 프로토타입의 첫 화면이다. 상품추천시스템의 프로토타입에 관한 설명과 온라인설문 절차에 대한 내용이다.

<Figure 3>은 추천에 사용할 기반정보를 토대로 추천결과를 보여주는 것이다. 제안추천시스템의 추천결과와 무작위 추천결과의 전체 화면레이아웃



<Figure 4> Survey Results of Satisfaction for the Recommendation

은 동일하며 순서는 임의적으로 나타난다.

온라인설문은 상품추천시스템 프로토타입의 추천결과에 대한 만족도를 5점 척도로 평가한다. 온라인설문대상자는 MSN 메신저사용자, 다음카페회원, P2P 사이트 회원이며, 총 173명이 설문에 참여하였다. <Figure 4>는 제안 추천시스템의 추천결과와 무작위 추천결과에 대한 만족도를 비교한 그래프이다.

<Figure 4>에서 좌측 그래프 'RECOMMEN'은 제안 추천시스템 모형의 추천결과에 대한 평가점수의 빈도수이고 우측 그래프 'RANDOM'은 무작위 추천결과에 대한 평가점수의 빈도수이다. X 축은 빈도수, Y축은 평가점수를 나타낸다. RECOMMEN' 그래프가 'RANDOM'에 비해 우측으로 치우쳐진 형태이다. 이 그림을 통해 무작위 추천결과에 비해 제안 추천시스템 모형의 추천결과가 높은 만족도를 보였음을 직관적으로 판단할 수 있다.

<Figure 4>의 결과에 대한 평균값 간의 차이가 통계적으로 유의한지를 확인하고자 대응표본 T검정(Paired Sample T-test)을 실시하였다. 검정결과 유의수준 1% 내에서 통계적으로 유의한 차이가 있음을 알 수 있었다. 그 결과는 <Table 7>과 같다.

<Table 7>의 결과는 본 연구에서 제안하는 상품추천시스템의 추천결과에 대한 이용자의 만족도와 임의추천방식의 추천결과에 대한 이용자의 만족도의 평균값의 차이가 통계적으로 유의함을 나타내고, 본 연구에서 제안하는 상품추첨시스템이 이용자의 만족도 측면에서 유용함을 나타낸다.

<Table 7> Statistical Test Results for the Survey of User Satisfaction

	T-value	Degree of Freedom	Significance
Proposed Recommender System-Random Selection	7.972	173	0.000

6. 결론

본 논문에서는 기존 상품추천시스템을 개선하기 위하여 각 상품군마다 로지스틱 회귀분석, 의사결정나무, 인공신경망의 분류모형을 만들고 이 결과를 다중모형조합기법의 Bumping과 Bagging으로 결합하여 특정 상품군을 선호한 고객을 분류하였으며, 상품 간의 연관규칙을 도출하였다. 제안한 상품추천시스템의 유용성을 확인하기 위해 실제 인터넷 사용자들에게 시스템 프로토타입을 제시하였으며

이를 통해 확인한 이용자의 만족도를 바탕으로 제안추천시스템의 추천결과가 무작위 추천결과에 비해 만족도가 높게 나타났으며 통계적으로 유의함으로 실제 적용가능성이 높음을 확인할 수 있었다.

본 연구는 많은 한계점을 가지고 있는데 첫째, 각 단일 분류모형의 결과 차이가 크지 않으므로 보다 정교한 분류모형을 개발할 필요가 있다. 둘째, 다중모형조합기법의 성과도 단일분류모형의 성과에 비해 유의하게 차이가 나지 않아서 보다 개선된 다중모형조합기법의 개발이 필요할 것으로 생각된다. 최근에 유전자 알고리즘과 같은 최적화 기법을 활용한 다중모형조합기법이 소개되고 있는데 이런 기법은 최적화 과정을 거쳐 단일모형의 단점을 보완할 수 있을 것으로 기대된다. 셋째, 연구의 결과가 보다 일반화되기 위해서는 보다 다양한 상품군에 대한 작용과 보다 많은 이용자에 대한 만족도 조사가 필요할 것으로 생각된다.

참고문헌

Adomavicius, G. and A. Tuzhilin, "Toward the next generation of recommender systems : a survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.6(2005), 734~749.

Ahn, H. C., I. Han, and K. Kim, "The Product Recommender System Combining Association Rules and Classification Models : The Case of G Internet Shopping Mall," *Information Systems Review*, Vol.8, No.1(2006), 181~201.

Breiman, L., "Heuristics of instability in model selection," *Technical Report*, Statistics Department, University of California at Berkeley, 1994.

Breiman, L., "Bagging predictors," *Machine Learning*, Vol.24, No.2(1996), 123~140.

Cho, Y. H. and J. K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert Systems with Applications*, Vol.26 (2004), 233~246.

Cho, Y. H., J. K. Kim, and S. H. Kim, "A personalized recommender system based on Web usage mining and decision tree induction," *Expert Systems with Applications*, Vol.23(2002), 329~342.

Cho, Y. H. and J. H. Bang, "Applying Centrality Analysis to Solve the Cold-Start and Sparsity Problems in Collaborative Filtering," *Journal of Intelligence and Information Systems*, Vol.17, No.3(2009), 183~199.

Cho, Y. H., S. K. Park, D. H. Ahn, and J. K. Kim, "Collaborative Recommendations using Adjusted Product Hierarchy : Methodology and Evaluation," *Journal of the Korean Operations Research and Management Science Society*, Vol.29, No.2(2004), 59~75.

Goldberg, D., D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, Vol.35, No.12(1992), 61~70.

Heskes, T., "Balancing between bagging and bumping," *Advances in Neural Information Processing Systems*, Cambridge, MIT Press, (1996), 466~472.

Kang, B., "Collaborative Filtering System using Self-Organizing Map for Web Personalization," *Journal of Intelligence and Information Systems*, Vol.9, No.3(2003), 117~135.

Kim, D. and B.-J. Yum, "Collaborative filtering based on iterative principal component analysis," *Expert Systems with Applications*, Vol.28,

No.4(2005), 823~830.

Kim, J. K., Y. H. Cho, W. J. Kim, J. R. Kim, and J. H. Suh, "A personalized recommendation procedure for Internet shopping support," *Electronic Commerce Research and Applications*, Vol.1(2002a), 301~313.

Kim, J. K., D. H. Ahn, and Y. H. Cho, "A Personalized Recommender System, WebCF-PT : A Collaborative Filtering using Web Mining and Product Taxonomy," *Asia Pacific Journal of Information Systems*, Vol.15, No.1(2005), 63~79.

Kim, J. K., D. H. Ahn, and Y. H. Cho, "Development of a personalized recommendation procedure based on data mining techniques for internet shopping malls," *Journal of Intelligence and Information Systems*, Vol.9, No.3(2003), 177~191.

Kim, J. K., J. H. Suh, D. H. Ahn, and Y. H. Cho, "A personalized recommendation methodology based on collaborative filtering," *Journal of Intelligence and Information Systems*, Vol.8, No.2(2002b), 139~157.

Kim, J. W., S. J. Bae, and H. J. Lee, "Sparsity Effect on Collaborative Filtering-based Personalized Recommendation," *Asia Pacific Journal of Information Systems*, Vol.14, No.2(2004), 131~149.

Kim, K. and H. Ahn, "Collaborative filtering with a user-item matrix reduction technique for recommender systems," *International Journal of Electronic Commerce*, Vol.16, No.1(2011), 107~128.

Kim, K. and. B. Kim, "Product Recommender System for Online Shopping Malls using Data Mining Techniques," *Journal of Intelligence and Information Systems*, Vol.11, No.1(2005), 191~205.

Konstan, j., B. Miller, D. Maltz, J. Herlocker, L. Gordon, and J. Riedl, "GroupLens : Applying Collaborative Filtering to Usenet News," *Communication of the ACM*, Vol.40(1997), 77~87.

Lee, Y. and S. Kwak, "A study on training ensembles of neural networks : a case of stock price prediction," *Journal of Intelligence and Information Systems*, Vol.5, No.1(1999), 95~101.

Park, J. H., Y. H. Cho, and J. K. Kim, "Social Network : A Novel Approach to New Customer Recommendations," *Journal of Intelligence and Information Systems*, Vol.15, No.1(2009), 123~140.

Pazzani, M. J., "A framework for collaborative, content-based and demographic filtering," *Artificial Intelligence Review*, Vol.13, No.5-6(1999), 393~408.

Resnick, P., N. Iacovou, M. Suchak, and P. Bergstrom, "GroupLens : An open architecture for collaborative filtering of netnews," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, (1994), 175~186.

Roh, T. H., K. J. Oh, and I. Han, "The collaborative filtering recommendation based on SOM cluster-indexing CBR," *Expert Systems with Applications*, Vol.25, No.3(2003), 413~423.

Sarwar, B. M., G. Karypis, J. A. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," *Proceedings of Conference on ACM*, (2000), 158~167.

Tibshirani, R. and K. Knight, "Model search and inference by bootstrap 'bumping'," *Technical Report*, University of Toronto, 1995.

Wu, K.-L., C. C. Aggarwal, and P. S. Yu, "Personalization with dynamic profiler," *Proceedings of the Third International Workshop on Advanced Issues of E-commerce and Web-based Information Systems*, (2001), 12~20.

Abstract

Product Recommender Systems using Multi-Model Ensemble Techniques

Yeonjeong Lee* · Kyoung-jae Kim**

Recent explosive increase of electronic commerce provides many advantageous purchase opportunities to customers. In this situation, customers who do not have enough knowledge about their purchases, may accept product recommendations. Product recommender systems automatically reflect user's preference and provide recommendation list to the users. Thus, product recommender system in online shopping store has been known as one of the most popular tools for one-to-one marketing. However, recommender systems which do not properly reflect user's preference cause user's disappointment and waste of time.

In this study, we propose a novel recommender system which uses data mining and multi-model ensemble techniques to enhance the recommendation performance through reflecting the precise user's preference. The research data is collected from the real-world online shopping store, which deals products from famous art galleries and museums in Korea. The data initially contain 5759 transaction data, but finally remain 3167 transaction data after deletion of null data. In this study, we transform the categorical variables into dummy variables and exclude outlier data.

The proposed model consists of two steps. The first step predicts customers who have high likelihood to purchase products in the online shopping store. In this step, we first use logistic regression, decision trees, and artificial neural networks to predict customers who have high likelihood to purchase products in each product group. We perform above data mining techniques using SAS E-Miner software. In this study, we partition datasets into two sets as modeling and validation sets for the logistic regression and decision trees. We also partition datasets into three sets as training, test, and validation sets for the artificial neural network model. The validation dataset is equal for the all experiments. Then we composite the results of each predictor using the multi-model ensemble techniques such as bagging and bumping. Bagging is the abbreviation of "Bootstrap Aggregation" and it composite outputs from several machine learning techniques for raising the performance and stability of prediction or classification.

* Graduate School of Management, Dongguk University_Seoul
** Corresponding Author: Kyoung-jae Kim
Business School, Dongguk University_Seoul
30, Pildong-ro 1gil, Jung-gu, Seoul 100-715, Korea
Tel: +82-2-2260-3324, Fax: +82-2-2260-3684, E-mail: kjkim@dongguk.edu

This technique is special form of the averaging method. Bumping is the abbreviation of “Bootstrap Umbrella of Model Parameter,” and it only considers the model which has the lowest error value. The results show that bumping outperforms bagging and the other predictors except for “Poster” product group. For the “Poster” product group, artificial neural network model performs better than the other models.

In the second step, we use the market basket analysis to extract association rules for co-purchased products. We can extract thirty one association rules according to values of Lift, Support, and Confidence measure. We set the minimum transaction frequency to support associations as 5%, maximum number of items in an association as 4, and minimum confidence for rule generation as 10%. This study also excludes the extracted association rules below 1 of lift value. We finally get fifteen association rules by excluding duplicate rules. Among the fifteen association rules, eleven rules contain association between products in “Office Supplies” product group, one rules include the association between “Office Supplies” and “Fashion” product groups, and other three rules contain association between “Office Supplies” and “Home Decoration” product groups. Finally, the proposed product recommender systems provides list of recommendations to the proper customers.

We test the usability of the proposed system by using prototype and real-world transaction and profile data. For this end, we construct the prototype system by using the ASP, Java Script and Microsoft Access. In addition, we survey about user satisfaction for the recommended product list from the proposed system and the randomly selected product lists. The participants for the survey are 173 persons who use MSN Messenger, Daum Café, and P2P services. We evaluate the user satisfaction using five-scale Likert measure. This study also performs “Paired Sample T-test” for the results of the survey. The results show that the proposed model outperforms the random selection model with 1% statistical significance level. It means that the users satisfied the recommended product list significantly. The results also show that the proposed system may be useful in real-world online shopping store.

Key Words : Product Recommender System, Multi-Model Ensemble Technique, Association Rules, Decision Tree, Artificial Neural Networks

저 자 소 개

이연정
동국대학교 일반대학원 경영정보학과 박사과정을 수료하고 현재 New Tang Dynasty Television에 재직 중이다. 동국대학교 정보관리학과에서 경영학사를 취득하고, 동교에서 경영정보학 석사를 취득하였다. 주요 관심분야는 데이터마이닝, 고객관계관리 등이다.

김경재
현재 동국대학교 경영대학 경영학부 교수로 재직 중이다. KAIST에서 경영정보시스템을 전공으로 박사학위를 취득하였으며, 경영학연구, 지능정보연구, Annals of Operations Research, Applied Intelligence, Applied Soft Computing, Asia Pacific Journal of Information Systems, Computers and Operations Research, Computers in Human Behavior, Expert Systems, Expert Systems with Applications, Information, Intelligent Data Analysis, International Journal of Electronic Commerce, Intelligent Systems in Accounting, Finance and Management, Neural Computing and Applications, Neurocomputing 등의 학술지에 논문을 게재하였다. 연구 관심분야는 고객관계관리, 데이터마이닝, 비즈니스 인텔리전스, 지식경영 등이다.