Journal of KIIT. Vol. 17, No. 10, pp. 115-120, Oct. 31, 2019. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2019.17.10.115

나이브 베이즈 분류를 활용한 부동산 추천 기법 연구

유 석 종*

A Study on Recommendation Method for Real Estate Using Naive Bayes Classification

Seok-Jong Yu*

요 약

본 연구에서는 나이브 베이즈 분류기를 활용하여 부동산 매물의 고유 특성에 기반한 추천 모델을 제안하고자 한다. 기존의 부동산 추천 시스템은 주로 거래량 또는 상승률 순위 리스트에 의존하고 있으며, 개별 매물의 속성을 통한 추천 형식은 찾아보기 어렵다. 매물의 건축년도, 지역, 면적, 등 속성 정보를 통해 향후 상승 전망을 나이브 베이즈 모델로 예측하는 것이 본 연구의 목표이다. 국토부의 아파트 실거래가 데이터를 이용하여 과거 아파트 매물의 각종 변화 지표를 추출하고 물건과 속성 지표간의 상관관계를 분석하여 추천 단계를 분류한다. 제안 모델 평가를 위해 국토부의 아파트 실거래가 데이터셋을 구축하고 분류 실험하였으며, 분류 정확도가 평균 85% 수준으로 측정되어 속성정보와 가격 상승률간 상관관계를 확인할 수 있었고, 이 분야에서 추천 모델로써 나이브 베이즈 분류기의 가능성을 확인할 수 있었다.

Abstract

In this study, a recommendation model using Naive Bayes classifier is proposed, which focuses on the intrinsic attributes of real estates. It is hard to find this kind of model that is distinguished from existing methods which mainly rely on the amounts of views or increased value ranking of real estates. The purpose of the study is how well Naive Bayes model predicts the rate of value rise in the future by attribute variables such as built year, location, area, story of a particular object. To accomplish this goal, it creates an initial dataset with actual contract value data of real estates from Ministry of Homeland, and then extract past guidances, analyze correlation with attributes, and classify to high probable ranking list level. From experimental results, 85% classification accuracy, it was possible to confirm both attribute-price co-relation and the possibility of the implementation model in this area.

Keywords

naive bayes classifier, recommender system, prediction, classification

* 숙명여자대학교 소프트웨어학부 교수
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Aug. 13, 2019, Revised: Oct. 16, 2019, Accepted: Oct. 19, 2019

· Corresponding Author: Seok-Jong Yu
Department of Software, Sookmyung Women's University, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

Ⅰ. 서 론

정보가 폭증하면서 많은 대상 중에 원하는 조건에 부합하는 것을 찾는 일이 흔해지고 있다. 이런 맥락으로 최근 추천 시스템에 관심과 연구가 증가하고 있으며, 아마존의 상품 추천과 넷플릭스의 영화 추천이 대표적인 성공 사례이다. 본 연구에는 사람들의 많은 관심이 집중되고 있는 부동산 영역에서 추천 시스템 주제를 적용하고자 한다. 부동산의 종류에는 주택, 아파트, 상가, 빌딩 등 다양한 것들이 존재하지만 실 데이터를 구하기 쉬운 아파트를 대상으로 하고자 한다. 부동산에 대한 추천 정보는 주로 전문 웹사이트, 앱, 신문 등의 매체를 통해서 획득하게 된다. 네이버 부동산은 가장 많이 활용되고 있는 사이트 이며 이밖에도 직방, 호갱노노, 등 다양한 부동산 추천 앱들이 이용되고 있다. 기존 시스템들의 추천 방법으로 가장 많이 사용되고 있는 방법은 사용자가 지역, 면적 등의 탐색 범위를 부여하고 이에 맞는 물건을 찾아 시간순, 가격순으로 정렬하여 보여준다. 다른 방법으로 그림 1과 같이, 최근 일정 기간 동안 부동산 물건에 대한 조회수, 실거래가 상승률, 등을 이용한 랭킹 기반 리스트 추천이다. 이러한 방법들도 원하는 물건을 탐색하는데 도움이 되지만 개별 물건의 특성이 이미 반영된 결과에 대한 추천이라고 볼 수 있다.

개별 물건에 내재되어 있는 특징, 예를 들어, 건축년도, 위치, 면적, 층, 브랜드 등과 같은 속성 정보를 반영하는 비조건 검색 기반 추천 방법은 부재한 상황이다.

그림 1. 조회 수 랭킹 (네이버 부동산)
Fig. 1. View ranking (land.naver.com)

예를 들어 건축년도, 위치, 면적, 브랜드와 같은 정보가 주어졌을 때 전체 물건에서 선호도가 어느 정도일지 예측할 수 있다면 소비자가 원하는 물건을 선택하는데 도움을 줄 수 있다.

본 연구에는 이러한 목표를 달성하기 위하여 나이브 베이즈 분류(Naive bayes classification) 방법을 활용한 연구를 진행하고자 하였다. 나이브 베이즈 분류는 텍스트 분석, 다중 클래스 분류(Multi-class classification), 스팸 필터링(Spam filtering), 감정분석(Sentiment analysis) 등에 사용된다[1][2]. 특히 알고리즘이 단순하고 성공률이 높으며 수행 속도가 빠른 것이 장점이다. 부동산 속성 변수는 독립적이기 때문에 나이브 베이즈 분류에 적용하기 적합하다. 이 분야에서 나이브 베이즈 모델을 활용한 추천 시스템 구현 사례는 확인되고 있지 않다.

본 논문의 구성은 다음과 같다. 2장은 관련 연구로 추천 시스템의 특징과 대표적인 추천 기법을 소개하고, 나이브 베이즈 모델에 대해 설명한다. 3장에서는 나이브 베이즈 모델을 사용한 추천 시스템 방법 및 설계에 대해 자세히 서술하고, 4장에서는 제안 방법의 성능 평가를 위한 실험 결과를 제시한다. 5장에서는 결론과 향후 연구 방향을 기술한다.

Ⅱ. 관련 연구

2.1 추천 시스템

추천 시스템(Recommender system)은 대량의 상품 중에서 소비자의 취향에 적합한 상품을 필터링하여 제안해 주는 시스템으로 소비자 및 상품의 속성 정보에 의존하는 내용 기반 방식과 유사 선호도를 갖는 이웃들의 상품 평가 이력 데이터에 기반하는 협업필터링(Collaborative filtering) 방식이 있다[3]. 내용 기반 추천은 구조적으로 사용자가 이미 경험한 것과 유사한 아이템이 주로 추천되는 포트폴리오(Portfolio) 현상이 불가피하게 발생되는 단점이 있다.

협업 필터링은 사용자의 과거 상품 평가 기록을 분석하여 선호도가 유사한 이웃을 탐색하고 미경험 상품에 대한 선호도를 예측하는 사회적 추천 방법이다. 따라서 협업 필터링의 추천의 질은 탐색된 이

웃에 달려 있으며, 뛰어난 추천 성능에도 불구하고 다음의 구조적 문제점이 발생한다[3-4]. 첫째, 평가 이력이 없는 초기 사용자나 신규 아이템은 추천이 어려운 초기 평가(Cold-start) 문제가 발생하며 개인화된 추천이 불가능하다. 두번째, 사용자의 평가 행렬이 희박(Sparsity)할 경우 연관된 이웃 탐색이 어려우며 추천의 질도 낮아진다. 끝으로 확장성(Scalability)은 사용자와 아이템의 수가 증가할수록 연산 시간이 비례하여 증가하며 실시간 추천을 어렵게 한다. 또한, 시간에 따라 변화하는 사용자의 선호도 문제를 위한 추천 다양성과 참신성(Novelty)에 대한 연구[5]가 진행되고 있다.

2.2 나이브 베이즈 분류

이메일을 중요도에 따라 스팸 메일과 햄메일로 구분하는 것은 빈번히 발생하는 대표적인 문서 분류의 작업이다[6]. 이미 분류된 문서들의 속성 정보(단어)를 통해 학습한 다음 미 분류된 문서들의 카테고리를 분류한다. 보통, 이러한 학습을 위해 문서에 포함된 단어들을 벡터로 수치화하는 작업이 필요하다. 이와 같은 문서 분류 방법으로 나이브 베이즈 분류기, KNN, SVM, 의사 결정 트리(Decision Tree), 인공신경망 등이 있다[1][7].

나이브 베이즈 모델은 단어들이 서로 독립된다고 가정할 때 특정 단어들을 포함하는 문서가 어느 범주에 속할 지 결정하는 모델로 속할 확률이 가장 큰 범주로 분류한다[8]. 베이즈 정리(Bayes' theorem)를 이용하여 개발된 이 알고리즘은 자료량이 많을수록 정확도가 높다. 사건 B가 일어났을 때 사건 A가 일어날 조건부 확률(P(A|B))과, 사건 A 확률(P(A)), 사건 B의 확률(P(B))을 가지고, 사건 A가 일어났을 때 사건 B의 조건부 확률(P(B|A))을 다음과 같이 계산할 수 있다.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (1)$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

나이브 베이즈 분류는 분류해야 할 객체 x의 속성 $x_1, x_2, \ldots x_n$이 서로 독립적이라 가정하고, 이 x를 카테고리 C_k로 분류하는 작업은 다음과 같이 계산한다.

$$x = [x_1\, x_2\, x_3 \cdots\, x_n] \quad (2)$$

$$arg_k \max P(C_k|x)$$

$$= arg_k \max P(x_1\, x_2\, x_3 \cdots\, x_n|C_k)P(C_k)$$

$$\approx arg_k \max \Pi_i P(x_i|C_k)P(C_k)$$

$P(x_1\, x_2\, x_3 \ldots x_n|C_k)$는 각 속성 변수들이 카테고리 C_k에 속할 확률의 곱을 의미하고, 이 확률이 가장 큰 C_k를 찾는 것이 나이브 베이즈 분류의 알고리즘이다. 나이브 베이즈 분류의 장점은 연산 속도가 빠르고 분류 정확도가 높다는 점이다. 또한 노이즈와 누락 데이터가 있어도 비교적 수행 성능이 좋으며 다중 분류 문제에도 잘 동작한다[9]. 나이브 베이즈 분류는 협업필터링과 결합하여 정보를 필터링하고 사용자가 특정 항목을 선호할지 여부를 예측하는 추천 시스템에 활용될 수 있다[10].

III. 나이브 베이즈 분류 기반 추천

3.1 나이브 베이즈 분류 기반 추천 개요

본 연구는 아파트의 속성 정보와 아파트 선호도와의 연관성을 나이브 베이즈 분류를 이용하여 확인하는데 목적이 있다. 일반적으로 부동산을 선호하는 이유는 개인마다 주관적이며 다양하게 나타난다. 본 연구에서는 객관적인 아파트 선호도 기준으로 일정 기간 동안 아파트의 실거래가 상승금액, 상승률, 거래량 데이터를 사용하였다. 일반적으로 상승률과 거래량 지표가 높을수록 많은 사람들이 선호하는 인기 아파트라고 판단할 수 있다. 본 연구 목적 달성을 위한 분류 및 검증은 다음의 단계로 진행된다.

(1) 실거래가 데이터 전처리 (2) 트레이닝/테스트 데이터셋 생성 (3) 아파트 선호도 순위 리스트 생성 (4) 추천 단계별 분할 (5) 테스트 데이터셋 나이브 베이즈 분류 (6) 추천 분류 정확도 평가

3.2 실거래 데이터셋 전처리 및 생성

본 연구에서 사용한 국토부 실거래가 데이터셋의 통계 정보는 표 1과 같다. 동일 단지의 아파트라고 하더라도 선호도가 다르기 때문에 면적에 따라 각각 구분하여 처리하였다. 국토부에서 제공하는 아파트 실거래가 데이터는 개별 매매 거래 건에 대한 시군구, 번지, 본번, 부번, 단지명, 전용면적(㎡), 계약년월, 계약일, 거래금액, 층, 건축년도, 도로명의 요소로 구성되어 있으며 이를 파이썬에서 불러들여 성분별로 분리한다. 국토부에서 제공하는 원본 csv 파일은 각 성분이 “,”로 구분되어 있는데 단지명안에 “,”가 또 사용되어 이 부분을 수작업으로 처리하는데 많이 시간이 소요된다.

표 1. 실거래가 데이터셋 정보
Table 1. Dataset information

거래 기간	2008 ~ 2019년
지역	서울특별시 25개구
총 매매 거래수	831,051 건
총 전월세 거래수	154,531 건
전체 단지의 수	7,038 단지
매매 거래된 단지 수 (평형별 구분)	26,329 개
전세 거래된 단지 수 (평형별 구분)	15,358 개

3.3 분류 시스템 설계

해당 데이터셋에 포함된 각 단지의 선호도에 따른 순위 리스트를 생성하기 위하여 최근 10년간(2008~2019년), 서울 지역 아파트의 실거래가 상승금액, 상승률, 거래량 지표에 가중치를 적용하고 그 총점을 계산하여 선호도 리스트를 생성하였다. 선호도 순위 리스트를 다중 추천 단계로 분할하고, 단계수 변화에 따른 나이브 베이즈 모델의 분류 정확도를 비교하고자 추천 단계를 2 단계 ~ 5 단계로 설정하였다. 예를 들어, 2단계는 추천(50)/비추천(50)으로 구분하고, 3단계는 추천(33)/중립(33)/비추천(33)과 같은 단계로 이루어진다. 표 2는 속성변수에 의한 분류 추천 결과의 예이다.

표 2. 속성 변수와 분류 추천 결과 예
Table 2. Attributes and recommendation examples

물건	속성 변수 (위치, 면적, 브랜드, 층, 건축년도)	분류 결과
1	용산구 청파동, 85, 래미안, 10, 2005	추천
2	마포구 공덕동, 79, 자이, 5, 1998	비추천
3	광진구 자양동, 120, 우성, 162003	중립
4	관악구 신림동, 95, 대림, 2, 2017	추천

IV. 실험 및 성능 평가

나이브 베이즈 모델을 사용하여 아파트 속성 변수에 의한 분류의 정확도를 측정하고자 하였다. 전체 아파트 단지(평형 구분)에 대해서 최근 10년간 실거래가 상승률, 상승금액, 거래량에 가중치를 곱해서 선호도 순위 리스트를 생성한다. 실거래가 데이터셋을 트레이닝 데이터셋과 테스트 데이터셋으로 분할하고 테스트 데이터셋에 포함된 개별 아파트를 분류기에 의해 분류한 카테고리와 실제 속한 추천 카테고리의 일치도를 계산하였다. 분류에 사용한 아파트의 속성 변수(loc, name, area, build, story)는 속성 변수의 개수와 분류 정확도를 확인하기 위하여 표 1과 같이 여러 조합으로 각각 수행하였다.

분류 결과는 다음 세가지로 구분된다. 분류 알고리즘에 의해 해당 아파트의 실제 카테고리에 올바르게 분류된 경우(Success), 인접 카테고리로 잘못 분류되었지만 실제 카테고리와의 오차가 허용치 이내인 경우(Close), 다른 카테고리로 분류가 잘못된 경우(Fail)이다. 속성 변수 조합에 대해 각각 분류 정확도를 비교한 결과 속성변수가 많을수록 정확도가 높게 나타나는 것을 알 수 있다. p, i, c는 선호도 평가 순위를 계산하는데 사용된 가중치로 각각 해당 기간 동안 아파트의 상승금액(Price), 상승률(Increase), 거래량(Contract)을 의미한다.

표 3은 아파트 속성 변수별 분류 정확도 실험 결과이다. 속성 변수의 수에 비례하여 분류 정확도가 상승하는 것을 알 수 있다.

표 4는 아파트의 속성 변수가 loc, name, area일 때, 선호도 리스트 생성에 사용된 성분 가중치에 따른 분류 정확도를 보여 준다. 상승액(p)의 비율이 높을수록 분류 정확도가 상승하는 것을 알 수 있다.

표 3. 속성 변수별 분류 정확도(%) (p=1.0, i=0.0, c=0.0)
Table 3. Classification accuracy by attribute variable

	[loc]	[loc, name]	[loc, name, area]	[loc, name, area, build, story]
Success	61.6	72.5	74.7	75.0
Close	14.9	11.2	11.1	11.2
Fail	23.5	16.3	14.2	13.8
Success + close	76.5	83.7	85.8	86.2

표 4. 구간 비율별 분류 정확도(%) (loc, name, area)
Table 4. Classification accuracy by weight

	p = 0.6 i = 0.1 c = 0.3	p = 0.7 i = 0.1 c = 0.2	p = 1.0 i = 0.0 c = 0.0
Success	73.9	74.1	74.7
Close	10.6	10.8	11.1
Fail	15.5	15.1	14.2
Success + close	84.5	84.9	85.8

그림 2는 추천 단계별 분류 정확도 그래프이다. x축은 2 ~ 5 추천 단계를 의미하며 단계가 많을수록 분류 정확도가 낮아지는 경향을 보여 주고 있다. 그림 3는 속성 변수의 수와 분류 정확도와의 상관관계를 보여주는 실험 결과이다. 속성 변수의 수가 1개일 경우 그림 2에 비하여 분류 정확도가 크게 떨어지는 것을 확인할 수 있다.

V. 결론 및 향후 과제

본 연구에서는 아파트의 속성 정보와 가격 상승률과의 상관관계를 확인하기 위하여 나이브 베이즈 모델을 활용한 실험 시스템을 구현하고 성능 평가 실험을 진행하였다. 나이브 베이즈 모델을 활용한 실험 결과, 평균 85% 정도의 분류 정확도를 보여주어 아파트의 속성 정보와 상승률 간에 높은 연관성이 있음을 알 수 있었고, 이와 같은 환경을 위한 추천 시스템 구현에 나이브 베이즈 모델의 가능성을 확인할 수 있었다. 해외 부동산 추천 관련 연구와 본 연구는 사용 데이터셋 및 목표 지표 등이 상이하여 직접적인 정량 비교는 불가능하였다.

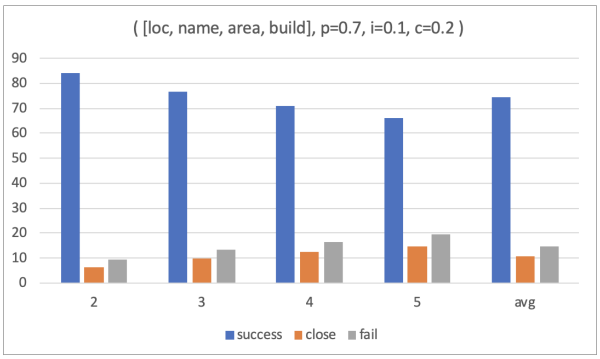
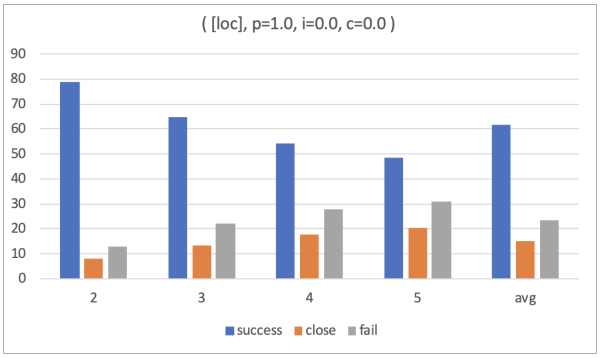
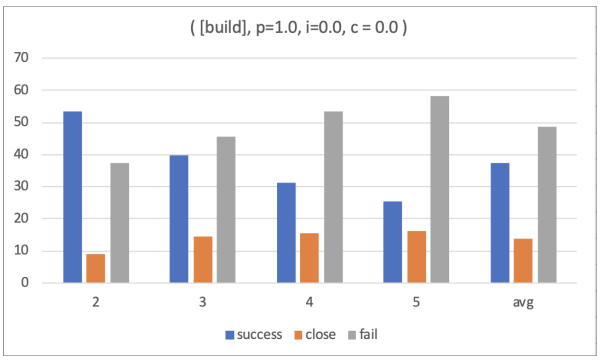


그림 2. 추천 단계별 분류 정확도
Fig. 2. Classification accuracy by level



(a) loc



(b) build

그림 3. 속성 변수별 분류 정확도
Fig. 3. Classification accuracy by attribute variable

본 연구의 목적은 단순히 분류 정확도 향상이 아니고 속성정보와 가격상승률간의 상관관계를 확인하는데 있으며 평균적으로 85%의 양호한 분류 결과를 얻었다. 실거래가 데이터 셋에서 제공되는 속성정보의 종류가 다양하지 않고 중복 정보가 많기 때문에 15%의 분류 오류가 발생했을 것으로 추측된다. 향후 연구에서 나이브 베이즈 모델과 다른 모델과의 하이브리드 방식을 통해 추천 및 분류 성능 향상을 기대해 볼 수 있을 것이다.

References

[1] Sanggi Lee, Byeong-Seop Lee, Byeong-Yong Bark, and Hyekyong Hwang, "A Study of Intelligent Recommendation System based on Naive Bayes Text Classification and Collaborative Filtering", Journal of Information Science Theory and Practice, Vol. 41, No. 4, pp. 227-249, Oct. 2010.

[2] J. Liu, Z. Tian, P. Liu, J. Jiang, and Z. Li, "An Approach of Semantic Web Service Classification Based on Naive Bayes", IEEE International Conference on Services Computing (SCC), San Francisco, CA, USA, pp. 356-362, Jun. 2016.

[3] G. Adomavicius and Y. Kwon, "Improving Aggregate Recommendation Diversity Using Ranking-based Techniques", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 5, pp. 896-911, May 2012.

[4] N. Lathia, S. Hailes, L. Capra, and X. Amatriain, "Temporal Diversity in Recommendation Systems", SIGIR, Geneva, Geneva, Switzerland, pp. 210-217, Jan. 2010.

[5] H. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem", Information Sciences, Vol. 178, No. 1, pp. 37-51, Jan. 2008.

[6] Seonghee Kim and Jae-Eun Eom, "A Study on the Document's Automatic Classification Using Machine Learning", Journal of Information Science Theory and Practice, Vol. 39, No. 4, pp. 47-66, Dec. 2008.

[7] Nam-Kuk Kim and Sang-Yong Lee, "Bayesian network based Music Recommendation System considering Multi-Criteria Decision Making", The Journal of digital policy & management, Vol. 11, No. 3, pp. 345-351, Mar. 2013.

[8] Alpaydin, Ethem, "Introduction to machine learning", Cambridge, MIT Press, 2014.

[9] Pedro Domingos and Michael Pazzani, "Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier", Proceedings of 13th International Conference on Machine Learning, pp. 105-122, 1996.

[10] Koji Miyahara and Michael J. Pazzani, "Improvement of Collaborative Filtering with the Simple Bayesian Classifier", Lecture notes in computer science 1886, 2000.

저자소개

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 컴퓨터과학과(이학사)
1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)
2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)
2005년 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수
관심분야 : 추천시스템, 협업필터링, 정보시각화