RFM 기법과 K-Means 알고리즘을 이용한 고객 분류

지현정*, 신경일, 신동일, 신동규
*세종대학교 컴퓨터공학과
e-mail : rrhak17@gce.sejong.ac.kr

A Study on Customer rating using RFM and K-Means

Hyunjung Ji, Gyeongil Shin, Dongil Shin, Dongkyoo Shin
Department of Computer Engineering, Sejong University, Korea

요 약

고객의 행동을 분석하기 위한 RFM(Recency, Frequency, Monetary)은 마케팅 분야에서 널리 쓰이고 있는 시장분석기법이다. 최근 축적되는 데이터가 많아지면서 이를 활용하기 위해 기계학습에 대한 관심이 증가하였다. 따라서 RFM 기법과 다양한 알고리즘을 결합하여 데이터를 분석하고자 하는 시도가 이루어지고 있다. 본 논문에서는 RFM 기법과 대표적인 클러스터링 알고리즘인 k-means 를 통하여 고객을 등급화 하는 방법에 대해 실험하였다. 기존의 실험에서는 k 값을 8 혹은 9 로 지정하는 사례가 많았다. 그러나 본 실험에서는 내부평가방법을 통해 데이터 셋에 대한 최적의 k 값을 구해보았고, 실험 결과 사용한 4 개의 데이터 셋에서 3 이라는 동일한 결과가 나왔다.

1. 서론

기업환경이 급격히 변화하는 현대 사회에서 기업들은 신규 고객을 유치하고, 기존 고객을 유지하기 위해 고객 관계 관리(CRM, Customer Relationship Management) 기법을 사용한다. CRM 은 지속적인 관계, 고객에 대한 개별 관리, 고객 정보 관리 등을 목적으로 널리 사용되고 있다 [1]. 또한 CRM 을 위한 데이터가 축적되면서, 이를 효율적으로 처리하기 위해 데이터 마이닝에 대한 관심이 급증하고 있다. 이러한 CRM 과 데이터 마이닝의 결합은 고객 관리에 있어서 단순한 통계 이상의 효과적인 결과를 낳을 것으로 기대된다.

RFM(Recency, Frequency, Monetary)은 고객의 행동을 분석하기 위해 널리 사용되는 마케팅 기법으로, 고객이 얼마나 최근(Recency)에 얼마나 자주(Frequency) 구매했는가, 그 구매의 규모(Monetary)는 얼마인가를 기준으로 고객의 가치를 분석한다. RFM 기법은 각 요소들을 통해 고객을 등급화 하게 된다 [2]. 기존의 많은 연구에서는 고객을 8 혹은 9 개의 클러스터로 나누는 것이 일반적이다 [2][5]. 그러나 본 논문에서는 RFM 기법과 대표적인 클러스터링 알고리즘인 k-means 를 이용하여 데이터에 따른 최적의 클러스터 개수를 구해보려 한다.

2. 관련연구

고객 관리에 관한 데이터 마이닝 기법을 적용한 사례는 다수 존재한다. Derya Birant 의 Data Mining Using RFM Anlaysis 백서에서는 RFM 기법을 통해 산출한 값을 토대로 클러스터링 알고리즘을 사용하여 고객의 등급을 분류하고 분류 알고리즘을 사용하여 신규고객의 정보로 고객의 등급을 예측한 뒤, 연관규칙 알고리즘을 통해 고객 개인에 맞는 상품을 추천해주는 과정을 소개하였다 [2]. 해당 백서에서 사용한 데이터 셋은 다양한 고객정보를 포함하고 있어 기존 고객에 대한 분석뿐만 아니라 신규 고객에 대한 예측까지 가능하였다.

2012 년 2 월, RFM 기법과 FP-tree 마이닝을 통한 개인 추천화 시스템을 개발한 사례가 있으며, 같은 해 6 월 k-means 기법을 이용하여 개인화 추천시스템을 개발하기도 하였다 [3][4]. 기존 추천시스템에 널리 사용되는 협업필터링은 사용자로부터 평점을 입력 받아야하는 명시적 방법이다. 이 연구에서는 이러한 명시적방법이 아닌 구매데이터를 통한 묵시적 방법이라는 점을 강조하였다.

2015 년에는 RFM 기반 SOM 을 이용한 매장관리 전략에 관한 연구가 있었다. 이 연구에서는 고객이 아닌 상품에 대한 RFM 을 도출하여 SOM 을 사용해 데이터를 군집화 한 후 각 군집 별 상품 관리 방법에 대한 전략을 제시하였다 [5]. 상품에 대한 RFM 을 도출하여 군집화 한 결과 주요 군집을 찾아낼 수 있었으며, 각 군집에 대한 분석을 통해 상품이 속한 군집의 특징에 따라 해당 상품의 배치 및 재고 수준 등을 관리할 수 있다.

3. 실험

3.1. 데이터 전처리

클러스터링 알고리즘을 실험하기에 앞서 데이터 전처리가 필요하다. 클러스터링 알고리즘에 입력으로 들어갈 속성은 RFM 기법에 따라 Recency, Frequency, Monetary 이다. 각 항목을 구하는 방법은 아래와 같다.

-Recency: 고객 별로 구매날짜 속성 중 가장 최근 항목만을 남기고 데이터를 제거한다. 구매날짜를 기준으로 정렬한 후 가장 오래된 날짜를 기준으로 각 고객의 구매 날짜와의 차를 계산하여 Recency 속성으로 한다.

-Frequency: 고객 별로 중복되지 않는 주문번호를 count 한 값으로 한다.

-Monetary: 고객 별로 주문번호를 중복제거한 후 최종 결제금액을 합한 값으로 한다.

RFM 속성을 산출 한 뒤 각 속성을 normalize 한다. 이는 각 속성별로 다양한 수의 범위를 가지고 있기 때문에 모두 0 에서 1 사이의 수치 값으로 대체시켜준 후 클러스터링 알고리즘을 실험하기 위함이다.

3.2 알고리즘 소개 및 실험방법

사용된 k-means 는 자율학습의 가장 대표적인 알고리즘으로 주어진 데이터를 k 개의 클러스터로 묶는다. 초기 k 개의 중심점을 잡고 각 데이터와 중심점 사이의 거리를 계산하여 해당 데이터에서 가장 가까운 클러스터를 찾아 데이터를 배당하게된다. k-means 알고리즘에서는 최적의 k 값이 존재하지 않는다. 데이터에 따라 k 값이 달라지게 되며 클러스터링 결과를 통해 최적의 k 를 찾아야 한다. 클러스터링 알고리즘을 평가하는 방법에는 내부평가와 외부평가가 있다. 외부평가는 라벨이 존재하는 데이터의 경우 라벨을 제외하고 클러스터링 알고리즘을 통해 결과를 라벨과 비교하여 알고리즘을 평가하는 방법이다. 반면 내부평가 방법을 사용하면 라벨이 없는 데이터 셋에 대한 클러스터링 결과를 평가할 수 있다. 내부평가는 클러스터 내의 밀집도와 클러스터 간의 분포를 통해 평가한다. 본 논문에서는 Silhouette coefficient 와 Calinski Harabaz 두가지의 내부평가 방법을 사용하여 데이터에 대한 최적의 k 를 구하려고 시도하였다.

다음은 Silhouette coefficient 의 계산 식이다. 이 수식에서 a 는 같은 클러스터 안에서 다른 모든 데이터들 간의 평균 거리이며, b 는 다른 클러스터의 개체들 사이의 평균 거리를 뜻한다. B 가 a 보다 크며 두 값 사이의 차가 클수록 s 는 1 에 가까워진다. S 값이 1 에 가까울수록 군집화가 잘 된 것이며 -1 에 가까울수록 잘 되지 않은 것이다. 즉 클러스터 내 밀집도가 크고 클러스터 간 거리가 멀수록 잘 구분되었다는 뜻이다 [6].

$$s = \frac{b-a}{max(a,b)}$$

다음은 Calinski Harabaz 의 계산 식이다. 수식에서 Wk 는 클러스터 내 분산행렬이며, Bk 는 클러스터 간 분산 행렬이다. s 값은 높을수록 클러스터가 잘 구분되었다는 뜻이며, Calinski Haravbaz 는 계산 속도가 빠르다는 장점이 있다 [6].

$$s(k) = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \times \frac{N-k}{k-1}$$

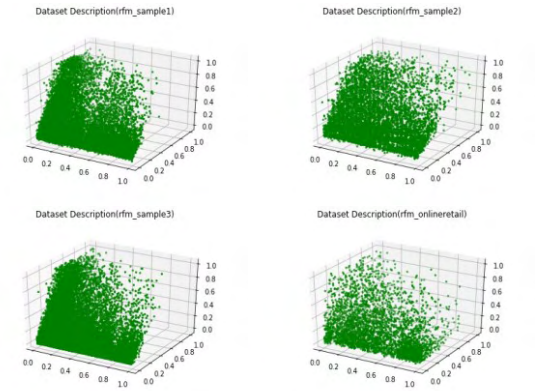
$$W_k = \sum_{q=1}^{k} \sum_{x \in C_q} (x - c_q)(x - c_q)^T$$

$$B_k = \sum_{q} n_q (c_q - c)(c_q - c)^T$$

실험은 python 의 표준 기계학습 라이브러리인 scikit-learn 을 통해 진행되었다 [7].

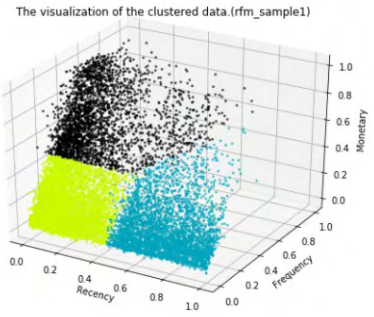
3.3. 실험결과

본 실험에 사용된 데이터 셋은 한 의류 쇼핑몰에서 고객 관리를 위해 수집한 데이터 셋 세가지(sample1, sample2, sample3 로 표기)와 UCI Repository 에서 제공받은 Online Retail Data Set(sample_onlineretail)을 사용하였다 [8]. 각 데이터 셋에 대한 전처리 방법은 본 논문 3 장과 같으며 아래 그림은 데이터 셋의 RFM 값을 normalize 한 후의 분포를 보여주는 그래프이다.



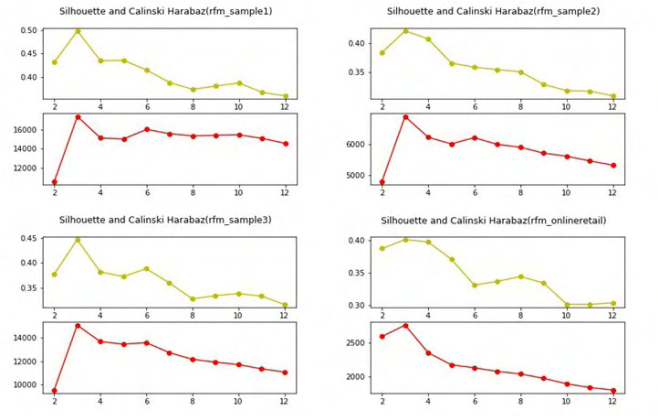
(그림 1) 사용한 네 가지 데이터 셋의 분포

각 데이터 셋에 대해 k 값을 2 개부터 12 개로 하여 k-means 알고리즘을 통해 실험하였다. 다음은 실험과정 중 한 예로 sample1 데이터에 대해 k 를 3 로 했을 때의 결과이다.



(그림 2) 데이터셋 sample1 을 3 개의 군집으로 군집화한 결과

다음은 각 데이터 셋에 대해 silhouette coefficient 와 calinski harabaz 두가지의 값을 산출한 후 그래프로 정리한 값이다.



(그림 3) 각 데이터 셋에 대한 내부평가 결과

실험 결과를 통해 사용된 데이터 셋에서 대부분 3 개의 클러스터로 분류하는 것이 가장 이상적이라는 결론을 얻었다. 4 개의 데이터 셋 모두에서 k 가 3 일 때 내부평가가 가장 좋게 나타난 것을 볼 수 있다.

다음은 4 개의 데이터 셋을 k 를 3 으로 하여 클러스터링한 결과, 각 클러스터의 평균을 정리한 표이다.

<표 1> 데이터셋 sample1 의 3 개의 군집에 대한 평균

	Average	*Cluster 1 (R ↑F ↓M ↓)*	*Cluster 2 (R ↓F ↓M ↓)*	*Cluster 3 (R ↓F ↑M ↑)*
Recency	0.3533	0.7223	0.1656	0.1786
Frequency	0.1523	0.0776	0.086	0.4041
Monetary	0.2887	0.1788	0.183	0.6763

<표 2> 데이터셋 sample1 의 클러스터링 결과

Cluster No.	*Instances(%)*
1	33
2	45
3	22

<표 3> 데이터셋 sample2 의 3 개의 군집에 대한 평균

	Average	*Cluster 1 (R ↑F ↓M ↑)*	*Cluster 2 (R ↓F ↑M ↑)*	*Cluster 3 (R ↓F ↓M ↓)*
Recency	0.4245	0.759	0.2845	0.2401
Frequency	0.217	0.1394	0.4754	0.1134
Monetary	0.3769	0.2514	0.708	0.2655

<표 4> 데이터셋 sample2 의 클러스터링 결과

Cluster No.	*Instances(%)*
1	33
2	26
3	40

<표 5> 데이터셋 sample3 의 3 개의 군집에 대한 평균

	Average	*Cluster 1 (R ↑F ↓M ↓)*	*Cluster 2 (R ↓F ↓M ↓)*	*Cluster 3 (R ↓F ↑M ↑)*
Recency	0.3451	0.7228	0.1959	0.1841
Frequency	0.1506	0.0736	0.0866	0.3931
Monetary	0.2721	0.1601	0.167	0.6516

<표 6> 데이터셋 sample3 의 클러스터링 결과

Cluster No.	*Instances(%)*
1	29
2	49
3	22

<표 7> 데이터셋 sample4 의 3 개의 군집에 대한 평균

	Average	*Cluster 1 (R ↓F ↓M ↓)*	*Cluster 2 (R ↓F ↑M ↑)*	*Cluster 3 (R ↑F ↓M ↓)*
Recency	0.4455	0.2285	0.2297	0.7868
Frequency	0.2234	0.1443	0.4878	0.1622
Monetary	0.2573	0.1801	0.5882	0.1581

<표 8> 데이터셋 sample4 의 클러스터링 결과

Cluster No.	*Instances(%)*
1	40
2	21
3	39

<표 1>에서 각 클러스터의 평균값과 전체 데이터의 평균값을 비교해 보았을 때 Cluster1 은 최근에 방문하였지만 자주방문하거나 많은 돈을 소비하지 않은 고객, Cluster2 는 최근에 방문하지도 자주 방문하지도 않으며 돈을 소비하지 않는 고객, Cluster3 은 최근에 방문하지는 않았지만 자주 방문하며 많은 돈을 소비하는 고객의 집합으로 분류되었다는 것을 알 수 있다. <표 3>에서 Cluster1 은 최근에 방문하였고, 많은 돈을 소비하였으나, 자주 방문하지 않는 고객의 집합이다. 실험 결과 <표 1>의 세개의 클러스터와 <표 3>의 Cluster1 로 네 가지(R↑F↓M↓, R↓F↓M↓, R↓F↑M↑, R↑F↓M↑) 이외의 클러스터 형태는 나타나지 않았다. 네 가지 클러스터 중 R↓F↑M↑와 R↓F↓M↓는 모든 데이터 셋에서 나타난 형태이며, R↑F↓M↓는 세개의 데이터 셋에서 그리고 R↑F↓M↑는 한 데이터 셋에서만 나타났다.

4. 결론

최근 CRM 과 데이터 마이닝을 결합하여 축적한 마케팅 데이터의 가치를 높이고자 하는 다양한 시도가 이루어지고 있다. 하지만 고객의 행동을 분석하기 이전에 해당 데이터를 제공한 쇼핑몰의 규모에 따라 R, F, M 값은 크게 차이 나게 된다. 따라서 모든 경우에 같은 척도를 적용하는 것은 합리적이지 못하다. 본 논문에서는 특정 쇼핑몰에서 제공한 데이터를 통하여 최적의 클러스터 개수를 예측해 보았다. 실험한 4 개의 데이터 셋의 경우 3 개의 클러스터로 고객을 분류하는 것이 가장 이상적이라는 결과를 얻었다. 이는 8 혹은 9 개의 클러스터로 고객을 분류하는 기존의 일반적인 방식과 상반되는 결과이다. 이는 단순히 k 의 값을 설정하지 않고 내부평가 값에 따라 절충적인 값을 선택해야 한다는 의미이다. 향후 연구에서는 산출한 등급과 함께 고객 개인별 맞춤 제품을 추천해주는 시스템을 고안할 예정이다.

ACKNOWLEDGMENT

이 논문은 2017 년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임(2017-0-00862. IOT 기반의 고객데이터 수집 및 자동

인식을 통한 빅데이터분석 클라우드 고객센싱 서비스 시스템 구축)

참고문헌

[1] Buttle, Francis, Customer relationship management: concepts and technologies, Routledge(2009)
[2] Birant, Derya. Data mining Using RFM Analysis, Knowledge-oriented applications in data mining, In Tech(2011)
[3] Young-Sung Cho, Mi-Sug Gu and Keun-Ho Ryu, 2012, Development of Personalized Recommendation System using RFM method and k-means Clustering, Journal of the Korea Society of Computer and Information , Vol. 17, No. 6, pp. 163~172.
[4] Young-Sung Cho and Ryu-Keun Ho, 2012, "Personalized Recommendation System using FP-tree Mining based on RFM," Journal of the Korea Society of Computer and Information , Vol. 17, No. 2, pp. 197~206.
[5] Yoon Jeong Jeong, Il Young Choi, Jae Kyeong Kim and Ju Choel Choi, 2015, "Strategy for Store Management Using SOM Based on RFM," Journal of Intelligence and Information Systems, Vol. 21, No. 2, pp. 93~112.
[6] scikit-learn.org. : http://scikit-learn.org/stable/modules/clustering.html
[7] scikit-learn.org. : http://scikit-learn.org/stable/index.html
[8] Online Retail dataset available: http://archive.ics.uci.edu/ml/datasets/online+retail Accessed on 7/26/2017