(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2019년11월26일
(11) 등록번호 10-2049166
(24) 등록일자 2019년11월20일

(51) 국제특허분류(Int. Cl.)
G06Q 30/02 (2012.01) *G06F 17/18* (2006.01)
(52) CPC특허분류
G06Q 30/0201 (2013.01)
G06F 17/18 (2013.01)
(21) 출원번호 10-2018-0024414
(22) 출원일자 2018년02월28일
심사청구일자 2018년02월28일
(65) 공개번호 10-2019-0103688
(43) 공개일자 2019년09월05일
(56) 선행기술조사문헌
KR101680055 B1*
KR1020090039882 A*
KR1020100056066 A*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
세종대학교산학협력단
서울특별시 광진구 능동로 209 (군자동, 세종대학교)
(72) 발명자
신동일
서울특별시 강남구 압구정로 347, 26동 1207호(압구정동, 한양아파트)
신동규
서울특별시 강남구 언주로 201, 1903호(도곡동, 에스케이리더스뷰)
지현정
제주특별자치도 제주시 정존11길 30, 노형동 302호(노형동, 동덕아파트)
(74) 대리인
양성보

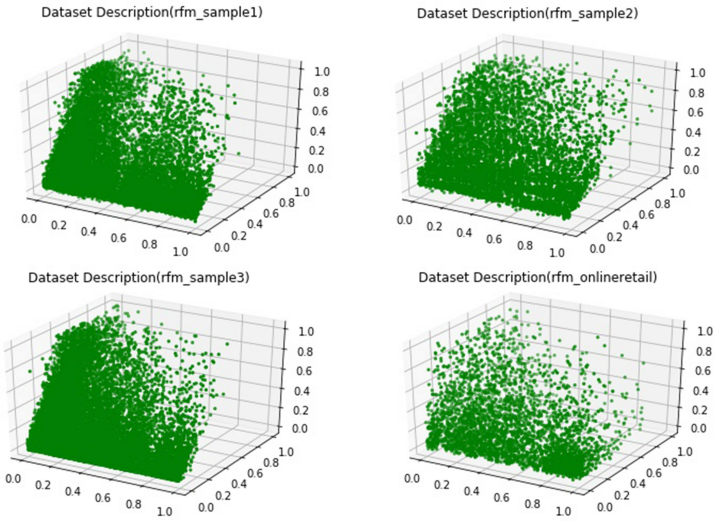
전체 청구항 수 : 총 10 항 심사관 : 홍경희

(54) 발명의 명칭 **RFM 기법과 K-Means 알고리즘을 이용한 고객 분류 방법 및 시스템**

(57) 요 약

일 실시예에 따른 고객 분류 시스템에 의하여 수행되는 고객 분류 방법은, 클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산하는 단계; 상기 계산된 거리에 기초하여 데이터와 가장 가까운 클러스터를 탐색하는 단계; 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계; 및 상기 적어도 하나 이상의 데이터 각각을 클러스터에 분류함에 따라 도출된 클러스터의 개수를 획득하는 단계를 포함할 수 있다.

대 표 도



이 발명을 지원한 국가연구개발사업

과제고유번호 1711055356

부처명 과학기술정보통신부

연구관리전문기관 정보통신기술진흥센터

연구사업명 ICT유망기술개발지원

연구과제명 IOT 기반의 고객데이터 수집 및 자동인식을 통한 빅데이터분석 클라우드 고객센싱 서비스 시스템 구축

기 여 율 1/1

주관기관 (주)소프트자이온

연구기간 2017.05.01 ~ 2018.04.30

공지예외적용 : 있음

명 세 서

청구범위

청구항 1

고객 분류 시스템에 의하여 수행되는 고객 분류 방법에 있어서,

클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산하는 단계;

상기 계산된 거리에 기초하여 데이터와 가장 가까운 클러스터를 탐색하는 단계;

상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계; 및

상기 적어도 하나 이상의 데이터 각각을 클러스터에 분류함에 따라 도출된 클러스터의 개수를 획득하는 단계

를 포함하고,

상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는,

상기 데이터와 가장 가까운 클러스터를 탐색하는 클러스터링 결과에 대한 내부 평가 및 외부 평가를 수행하고, 상기 데이터에 라벨이 존재할 경우 상기 라벨을 제외하고, 상기 클러스터링 결과를 상기 라벨과 비교하여 외부 평가를 수행하고, 상기 데이터에 라벨이 존재하지 않을 경우, 상기 라벨이 없는 데이터에 대한 클러스터링 결과를 평가하는 내부 평가를 수행하는 단계

를 포함하고,

상기 내부 평가는, 상기 클러스터 내의 밀집도와 클러스터 간의 분포를 통해 평가되고, Silhouette coefficient와 Calinski Harabaz를 포함하는 내부 평가 방법을 사용하여 상기 데이터에 대응하는 클러스터의 개수가 도출되는

고객 분류 방법.

청구항 2

제1항에 있어서,

상기 데이터를 R 속성, F 속성 및 M 속성으로 변환하는 전처리 과정을 수행하는 단계

를 더 포함하고,

상기 데이터를 R 속성, F 속성 및 M 속성으로 변환하는 전처리 과정을 수행하는 단계는,

상기 데이터에 대하여 상기 R속성, F 속성 및 M 속성을 포함하는 RFM 속성으로 산출함에 따라 각각의 속성을 정규화하는 단계

를 포함하는 고객 분류 방법.

청구항 3

삭제

청구항 4

삭제

청구항 5

삭제

청구항 6

삭제

청구항 7

제1항에 있어서,

상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는,

수학식 1의 계산을 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 단계

를 포함하고,

수학식 1:

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

a는 동일한 클러스터 내에서 다른 모든 데이터들 간의 평균 거리이고,

b는 상기 동일한 클러스터 이외의 다른 클러스터의 개체들 사이의 평균 거리를 의미하는, 고객 분류 방법.

청구항 8

제1항에 있어서,

상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는,

수학식 2를 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 단계

를 포함하고,

수학식 2:

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

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

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

W_k는 클러스터 내의 분산 행렬이고, B_k는 클러스터 간 분산 행렬을 의미하는, 고객 분류 방법.

청구항 9

제1항에 있어서,

상기 클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산하는 단계는,

상기 적어도 하나 이상의 데이터 각각을 기 설정된 개수의 클러스터로 분류하고, 상기 기 설정된 개수의 클러스터에 분류된 적어도 하나 이상의 데이터 각각과 상기 중심점과의 거리를 계산하는 단계

를 포함하는 고객 분류 방법.

청구항 10

고객 분류 시스템에 의하여 수행되는 고객 분류 시스템에 있어서,

클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산하는 계산부;

상기 계산된 거리에 기초하여 데이터와 가장 가까운 클러스터를 탐색하는 탐색부;

상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 분류부; 및

상기 적어도 하나 이상의 데이터 각각을 클러스터에 분류함에 따라 도출된 클러스터의 개수를 획득하는 획득부

를 포함하고,

상기 분류부는,

상기 데이터와 가장 가까운 클러스터를 탐색하는 클러스터링 결과에 대한 내부 평가 및 외부 평가를 수행하고, 상기 데이터에 라벨이 존재할 경우 상기 라벨을 제외하고, 상기 클러스터링 결과를 상기 라벨과 비교하여 외부 평가를 수행하고, 상기 데이터에 라벨이 존재하지 않을 경우, 상기 라벨이 없는 데이터에 대한 클러스터링 결과를 평가하는 내부 평가를 수행하는 것을 포함하고,

상기 내부 평가는, 상기 클러스터 내의 밀집도와 클러스터 간의 분포를 통해 평가되고, Silhouette coefficient와 Calinski Harabaz를 포함하는 내부 평가 방법을 사용하여 상기 데이터에 대응하는 클러스터의 개수가 도출되는

고객 분류 시스템.

청구항 11

제10항에 있어서,

상기 데이터를 R 속성, F 속성 및 M 속성으로 변환하는 전처리 과정을 수행하는 전처리부

를 더 포함하고,

상기 전처리부는,

상기 데이터에 대하여 상기 R속성, F 속성 및 M 속성을 포함하는 RFM 속성으로 산출함에 따라 각각의 속성을 정규화하는

것을 특징으로 하는 고객 분류 시스템.

청구항 12

삭제

청구항 13

삭제

청구항 14

삭제

청구항 15

제10항에 있어서,

상기 분류부는,

수학식 1의 계산을 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 것을 포함하고,

수학식 1:

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

a는 동일한 클러스터 내에서 다른 모든 데이터들 간의 평균 거리이고,

b는 상기 동일한 클러스터 이외의 다른 클러스터의 개체들 사이의 평균 거리를 의미하는, 고객 분류 시스템.

청구항 16

제10항에 있어서,

상기 분류부는,

수학식 2를 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 것을 포함하고,

수학식 2:

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

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

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

W_k는 클러스터 내의 분산 행렬이고, B_k는 클러스터 간 분산 행렬을 의미하는, 고객 분류 시스템.

청구항 17

제10항에 있어서,

상기 계산부는,

상기 적어도 하나 이상의 데이터 각각을 기 설정된 개수의 클러스터로 분류하고, 상기 기 설정된 개수의 클러스터에 분류된 적어도 하나 이상의 데이터 각각과 상기 중심점과의 거리를 계산하는

것을 특징으로 하는 고객 분류 시스템.

발명의 설명

기 술 분 야

[0001] 아래의 설명은 고객을 분류하는 기술에 관한 것이다.

배 경 기 술

[0002] 기업환경이 급격히 변화하는 현대 사회에서 기업들은 신규 고객을 유치하고, 기존 고객을 유지하기 위해 고객 관계 관리(CRM, Customer Relationship Management) 기법을 사용한다. CRM은 지속적인 관계, 고객에 대한 개별 관리, 고객 정보 관리 등을 목적으로 널리 사용되고 있다. 또한 CRM을 위한 데이터가 축적되면서, 이를 효율적으로 처리하기 위해 데이터 마이닝에 대한 관심이 급증하고 있다. 이러한 CRM과 데이터 마이닝의 결합은 고객 관리에 있어서 단순한 통계 이상의 효과적인 결과를 낳을 것으로 기대된다.

[0003] 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. 기존의 많은 연구에서는 고객을 8 혹은 9개의 클러스터로 나누는 것이 일반적이다.

[0004] 더 나아가, 이러한 데이터에 대한 최적의 클러스터 개수를 도출하여 보다 고객을 정확하게 분류하는 기술이 제안될 필요가 있다.

발명의 내용

해결하려는 과제

[0005] RFM 기법과 대표적인 클러스터링 알고리즘인 k-means를 이용하여 데이터에 따른 최적의 클러스터 개수를 도출하

는 고객 분류 시스템 및 방법을 제공할 수 있다.

과제의 해결 수단

[0006] 고객 분류 시스템에 의하여 수행되는 고객 분류 방법은, 클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산하는 단계; 상기 계산된 거리에 기초하여 데이터와 가장 가까운 클러스터를 탐색하는 단계; 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계; 및 상기 적어도 하나 이상의 데이터 각각을 클러스터에 분류함에 따라 도출된 클러스터의 개수를 획득하는 단계를 포함할 수 있다.

[0007] 상기 고객 분류 방법은, 상기 데이터를 R 속성, F 속성 및 M 속성으로 변환하는 전처리 과정을 수행하는 단계를 더 포함하고, 상기 데이터를 R 속성, F 속성 및 M 속성으로 변환하는 전처리 과정을 수행하는 단계는, 상기 데이터에 대하여 상기 R속성, F 속성 및 M 속성을 포함하는 RFM 속성으로 산출함에 따라 각각의 속성을 정규화하는 단계를 포함할 수 있다.

[0008] 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는, 상기 데이터와 가장 가까운 클러스터를 탐색하는 클러스터링 결과에 대한 내부 평가 및 외부 평가를 수행하는 단계를 포함할 수 있다.

[0009] 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는, 상기 데이터에 라벨이 존재할 경우 상기 라벨을 제외하고, 상기 클러스터링 결과를 상기 라벨과 비교하여 외부 평가를 수행하는 단계를 포함할 수 있다.

[0010] 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는, 상기 데이터에 라벨이 존재하지 않을 경우, 상기 라벨이 없는 데이터에 대한 클러스터링 결과를 평가하는 내부 평가를 수행하는 단계를 포함할 수 있다.

[0011] 상기 내부 평가는, 상기 클러스터 내의 밀집도와 클러스터 간의 분포를 통해 평가되고, Silhouette coefficient와 Calinski Harabaz를 포함하는 두 가지의 내부 평가 방법을 사용하여 상기 데이터에 대응하는 클러스터의 개수가 도출될 수 있다.

[0012] 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는, 수학식

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

1( )의 계산을 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 단계를 포함하고, a는 동일한 클러스터 내에서 다른 모든 데이터들 간의 평균 거리이고, b는 상기 동일한 클러스터 이외의 다른 클러스터의 개체들 사이의 평균 거리를 의미할 수 있다.

[0013] 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 단계는, 수학식

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

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

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

2( )를 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 단계를 포함하고, W_k는 클러스터 내의 분산 행렬이고, B_k는 클러스터 간 분산 행렬을 의미할 수 있다.

[0014] 상기 클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산하는 단계는, 상기 적어도 하나 이상의 데이터 각각을 기 설정된 개수의 클러스터로 분류하고, 상기 기 설정된 개수의 클러스터에 분류된 적어도 하나 이상의 데이터 각각과 상기 중심점과의 거리를 계산하는 단계를 포함할 수 있다.

[0015] 고객 분류 시스템에 의하여 수행되는 고객 분류 시스템은, 클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산하는 계산부; 상기 계산된 거리에 기초하여 데이터와 가장 가까운 클러

스터를 탐색하는 탐색부; 상기 탐색된 클러스터에 상기 적어도 하나 이상의 데이터 각각을 분류하는 분류부; 및 상기 적어도 하나 이상의 데이터 각각을 클러스터에 분류함에 따라 도출된 클러스터의 개수를 획득하는 획득부를 포함할 수 있다.

[0016] 상기 고객 분류 시스템은, 상기 데이터를 R 속성, F 속성 및 M 속성으로 변환하는 전처리 과정을 수행하는 전처리부를 더 포함하고, 상기 전처리부는, 상기 데이터에 대하여 상기 R속성, F 속성 및 M 속성을 포함하는 RFM 속성으로 산출함에 따라 각각의 속성을 정규화할 수 있다.

[0017] 상기 분류부는, 상기 데이터와 가장 가까운 클러스터를 탐색하는 클러스터링 결과에 대한 내부 평가 및 외부 평가를 수행할 수 있다.

[0018] 상기 분류부는, 상기 데이터에 라벨이 존재할 경우 상기 라벨을 제외하고, 상기 클러스터링 결과를 상기 라벨과 비교하여 외부 평가를 수행할 수 있다.

[0019] 상기 분류부는, 상기 데이터에 라벨이 존재하지 않을 경우, 상기 라벨이 없는 데이터에 대한 클러스터링 결과를 평가하는 내부 평가를 수행할 수 있다.

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

[0020] 상기 분류부는, 수학식 1( )의 계산을 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 것을 포함하고, a는 동일한 클러스터 내에서 다른 모든 데이터들 간의 평균 거리이고, b는 상기 동일한 클러스터 이외의 다른 클러스터의 개체들 사이의 평균 거리를 의미할 수 있다.

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

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

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

[0021] 상기 분류부는, 수학식 2 ( )를 통하여 상기 클러스터링 결과에 대한 내부 평가를 수행하는 것을 포함하고, W_k는 클러스터 내의 분산 행렬이고, B_k는 클러스터 간 분산 행렬을 의미할 수 있다.

[0022] 상기 계산부는, 상기 적어도 하나 이상의 데이터 각각을 기 설정된 개수의 클러스터로 분류하고, 상기 기 설정된 개수의 클러스터에 분류된 적어도 하나 이상의 데이터 각각과 상기 중심점과의 거리를 계산할 수 있다.

발명의 효과

[0023] 일 실시예에 따른 고객 분류 시스템은 최적의 클러스터의 개수를 도출함으로써 고객을 분류/고객의 행동을 분석할 수 있다.

[0024] 일 실시예에 따른 고객 분류 시스템은 기존의 고객 및 신규 고객에 대한 고객 관리가 가능하다.

도면의 간단한 설명

[0025] 도 1은 일 실시예에 따른 고객 분류 시스템의 구성을 설명하기 위한 블록도이다.

도 2는 일 실시예에 따른 고객 분류 시스템의 고객 분류 방법을 설명하기 위한 흐름도이다.

도 3은 일 실시예에 따른 고객 분류 시스템에서 사용된 네 가지 데이터 셋의 분포를 나타낸 것이다.

도 4는 일 실시예에 따른 고객 분류 시스템에서 데이터 셋을 기 설정된 개수의 클러스터로 클러스터링한 결과를 나타낸 예이다.

도 5는 일 실시예에 따른 고객 분류 시스템에서 각 데이터 셋에 대한 내부 평가 결과를 나타낸 것이다.

발명을 실시하기 위한 구체적인 내용

[0026] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

[0027] 실시예에서는 마케팅 분야에서 널리 쓰이고 있는 시장분석 기법 중 하나인 RFM을 통하여 고객의 행동을 분석하기 위하여 다양한 알고리즘을 결합하여 데이터를 분석할 수 있다. 최근 축적되는 데이터가 증가됨에 따라 기계학습에 대한 관심이 증가하였다. 이에 따라 RFM 기법과 k-means 알고리즘을 통하여 고객을 등급화하는 방법에 대하여 설명하기로 한다.

[0028] 도 1은 일 실시예에 따른 고객 분류 시스템의 구성을 설명하기 위한 블록도이고, 도 2는 일 실시예에 따른 고객 분류 시스템의 고객 분류 방법을 설명하기 위한 흐름도이다.

[0029] 고객 분류 시스템(100)은 계산부(110), 탐색부(120), 분류부(130) 및 획득부(140)를 포함할 수 있다. 이러한 구성요소들은 고객 분류 시스템(100)에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다. 구성요소들은 도 2의 고객 분류 방법이 포함하는 단계들(210 내지 240)을 수행하도록 고객 분류 시스템(100)을 제어할 수 있다. 이때, 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다.

[0030] 프로세서는 데이터 분석 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 고객 분류 시스템(100)에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 서버를 제어할 수 있다. 이때, 프로세서 및 프로세서가 포함하는 계산부(110), 탐색부(120), 분류부(130) 및 획득부(140) 각각은 메모리에 로딩된 프로그램 코드 중 대응하는 부분의 명령을 실행하여 이후 단계들(210 내지 240)을 실행하기 위한 프로세서의 서로 다른 기능적 표현들일 수 있다.

[0031] 단계(210)에서 계산부(110)는 클러스터의 각각에 설정된 중심점과 적어도 하나 이상의 데이터 각각의 사이의 거리를 계산할 수 있다. 계산부(110)는 적어도 하나 이상의 데이터 각각을 기 설정된 개수의 클러스터로 분류하고, 기 설정된 개수의 클러스터에 분류된 적어도 하나 이상의 데이터 각각과 중심점과의 거리를 계산할 수 있다. 이때, 사전에 사용자에 의하여 클러스터의 개수가 설정될 수 있다.

[0032] 단계(220)에서 탐색부(120)는 계산된 거리에 기초하여 데이터와 가장 가까운 클러스터를 검색할 수 있다.

[0033] 단계(230)에서 분류부(130)는 탐색된 클러스터에 적어도 하나 이상의 데이터 각각을 분류할 수 있다. 분류부(130)는 데이터와 가장 가까운 클러스터를 탐색하는 클러스터링 결과에 대한 내부 평가 및 외부 평가를 수행할 수 있다. 분류부(130)는 데이터에 라벨이 존재할 경우 라벨을 제외하고, 클러스터링 결과를 라벨과 비교하여 외부 평가를 수행할 수 있다. 분류부(130)는 데이터에 라벨이 존재하지 않을 경우, 라벨이 없는 데이터에 대한 클러스터링 결과를 평가하는 내부 평가를 수행할 수 있다. 이때, 내부 평가는, 클러스터 내의 밀집도와 클러스터 간의 분포를 통해 평가되고, Silhouette coefficient와 Calinski Harabaz를 포함하는 두 가지의 내부 평가 방법을 사용하여 데이터에 대응하는 클러스터의 개수가 도출될 수 있다.

[0034] 단계(240)에서 획득부(140)는 적어도 하나 이상의 데이터 각각을 클러스터에 분류함에 따라 도출된 클러스터의 개수를 획득할 수 있다.

[0035] 도 3은 일 실시예에 따른 고객 분류 시스템에서 사용된 네 가지 데이터 셋의 분포를 나타낸 것이다.

[0036] 고객 분류 시스템은 클러스터링을 수행하기 전에, 데이터 전처리 과정을 수행할 수 있다. 고객 분류 시스템은 RFM 기법에 따라 데이터에 대한 각각의 속성을 도출할 수 있다. 다시 말해서, 데이터를 R(Recency) 속성, F(Frequency) 속성 및 M(Monetary) 속성으로 변환하는 전처리 과정을 수행할 수 있다.

[0037] 이때, R 속성은, 고객 별로 구매날짜 중 가장 최근 항목만을 남기고 데이터를 제거한다. 구매날짜를 기준으로 데이터를 정렬하고 가장 오래된 날짜를 기준으로 각 고객의 구매 날짜와의 차를 계산한다. F 속성은, 고객별로 중복되지 않는 주문번호를 카운트한 값으로 한다. M 속성은, 고객별로 주문번호를 중복 제거한 최종 결제금액을 합산한 값으로 한다. 이와 같이, R 속성, F 속성 및 M 속성을 포함하는 RFM 속성을 산출한 뒤 각 속성을 정

규화할 수 잇다. 이는, 각 속성별로 다양한 수의 범위를 가지고 있기 때문에 모두 0 내지 1 사이의 수치 값으로 대체시킨 후 클러스터링을 수행하기 위함이다.

[0038] 고객 분류 시스템은 적어도 하나 이상의 데이터 각각을 기 설정된 개수의 클러스터로 분류할 수 있다. 고객 분류 시스템은 기 설정된 개수의 클러스터에 분류된 적어도 하나 이상의 데이터 각각과 중심점과의 거리를 계산할 수 있다. 고객 분류 시스템은 계산된 거리에 기초하여 데이터와 가장 가까운 클러스터를 탐색할 수 있다. 실시예에서는 자율학습의 가장 대표적인 알고리즘인 K-means를 사용하여 주어진 데이터를 k(k는 자연수)개의 클러스터로 묶을 수 있다. 초기 k개의 중심점을 잡고 각 데이터와 중심점 사이의 거리를 계산하여 데이터에서 가장 가까운 클러스터를 탐색하여 데이터를 가장 가까운 클러스터에 배당할 수 있다. K-means 알고리즘에서는 최적의 k값이 존재하지 않는다. 데이터에 따라 k값이 달라지게 되며, 클러스터링을 수행함에 따른 클러스터링 결과를 통하여 최적을 k를 탐색하여야 한다.

[0039] 고객 분류 시스템은 클러스터링 결과를 평가하기 위하여 내부 평가 및 외부 평가를 수행할 수 있다. 고객 분류 시스템은 데이터에 라벨이 존재할 경우 라벨을 제외하고, 클러스터링 결과를 라벨과 비교하여 외부 평가를 수행할 수 있다. 또는, 고객 분류 시스템은 데이터에 라벨이 존재하지 않을 경우, 라벨이 없는 데이터에 대한 클러스터링 결과를 평가하는 내부 평가를 수행할 수 있다. 이때, 내부 평가는, 상기 클러스터 내의 밀집도와 클러스터 간의 분포를 통해 평가되고, Silhouette coefficient와 Calinski Harabaz를 포함하는 두 가지의 내부 평가 방법을 사용하여 데이터에 대응하는 클러스터의 개수가 도출될 수 있다. 이러한 두 가지의 내부 평가 방법을 통하여 데이터에 대한 최적의 k가 도출될 수 있다.

[0040] 고객 분류 시스템은 수학식 1의 계산을 통하여 클러스터링 결과에 대한 내부 평가를 수행할 수 있다. 수학식 1은 Silhouette coefficient의 계산식이다.

[0041] 수학식 1:

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

[0042]

[0043] a는 동일한 클러스터 내에서 다른 모든 데이터들 간의 평균 거리이고, b는 동일한 클러스터 이외의 다른 클러스터의 개체들 사이의 평균 거리를 의미할 수 있다. b가 a보다 크며 두 값 사이의 차가 클수록 s는 1에 가까워진다. s값이 1에 가까울수록 군집화(클러스터링)가 잘 된 것이며, -1에 가까울수록 군집화가 잘되지 않은 것을 의미한다. 다시 말해서, 클러스터 내의 밀집도가 크고 클러스터 간 거리가 멀수록 잘 구분되었다는 것을 의미한다.

[0044] 고객 분류 시스템은 수학식 2를 통하여 클러스터링 결과에 대한 내부 평가를 수행할 수 있다. 수학식 2는 Calinski Harabaz의 계산식이다.

[0045] 수학식 2:

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

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

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

[0046]

[0047] 이때, W_k는 클러스터 내의 분산 행렬이고, B_k는 클러스터 간 분산 행렬을 의미할 수 있다. s값이 높을수록 클러스터가 잘 구분되었다는 뜻이다. Calinski Harabaz의 계산식은 계산 속도가 빠르다는 장점이 있다. 일례로, python의 표준 기계학습 라이브러리인 scikit-learn을 통하여 프로세스를 수행할 수 있다.

[0048] 도 3을 참고하면, 각 데이터 셋에 대한 처리 방법이 수행한 것을 나타낸 도면이다. 도 3은 데이터 셋의 RFM 값을 정규화한 후의 분포를 나타낸 그래프이다. 일례로, 데이터 셋이 의류 쇼핑몰에서 고객 관리를 위하여 수집된 데이터 셋 세가지(sample 1, sample 2, sample 3로 기재)와 UCI Repository에서 제공받은 Online Retail

Data Set(sample_onlineretail)이 사용될 수 있다. 각 데이터 셋에 대하여 k값을 2부터 12개로 하여 k-means 알고리즘을 통하여 클러스터링을 수행할 수 있다. 예를 들면, 도 4를 참고하면, sample 1 데이터에 대하여 k를 3으로 하였을 때의 결과를 나타낸 것이다. 다시 말해서, 데이터 셋 sample 1을 3개의 클러스터로 클러스터링한 클러스터링 결과를 나타낸 예이다.

[0049] 각 데이터 셋에 대하여 Silhouette coefficient와 Calinski Harabaz 두 가지의 값을 산출한 후 그래프로 나타낼 수 있다. 도 5를 참고하면, 각 데이터 셋에 대한 내부 평가 결과를 나타낸 것이다. 고객 분류 시스템은 클러스터링 결과를 통하여 사용된 데이터 셋에서, 예를 들면, 3개의 클러스터로 분류하는 것이 가장 이상적임을 판단할 수 있다. 4개의 데이터 셋 모두에서 k가 3일 때 내부 평가가 가장 좋게 나타난 것을 알 수 있다. 예를 들면, 3개의 클러스터는 최근에 방문하였지만 자주 방문하거나 많은 돈을 소비하지 않은 고객, 최근에 방문하지도 자주 방문하지도 않으며 돈을 소비하지 않는 고객, 최근에 방문하지는 않았지만 자주 방문하며 많은 돈을 소비하는 고객의 집합으로 분류될 수 있다. 이에 따라 종래에 정해진 8개 혹은 9개의 클러스터로 고객을 분류하는 일반적인 방식과 달리, 단순히 k값을 설정하지 않고 내부 평가 값에 따라 절충적인 값을 선택함으로써 고객을 분류/고객의 행동을 분석할 뿐만 아니라 고객 개인별 맞춤 제품을 추천할 수 있다. 다시 말해서, 고객 분류 시스템은 고객의 데이터에 따라 최적의 k값을 도출함으로써 고객을 등급화할 수 있다. 또한, CRM과 데이터 마이닝을 결합하여 마케팅 데이터의 가치를 높일 수 있다. 또한, 데이터를 제공한 업체의 규모에 따라 차이가 발생하는 속성값을 고려하기 때문에 동일한 척도를 적용했었던 불합리한 점을 극복하고, 합리적인 척도를 제공할 수 있다.

[0050] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 콘트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0051] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

[0052] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

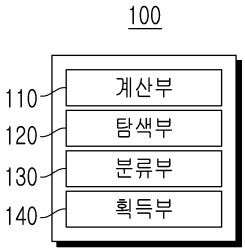
[0053] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태

로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

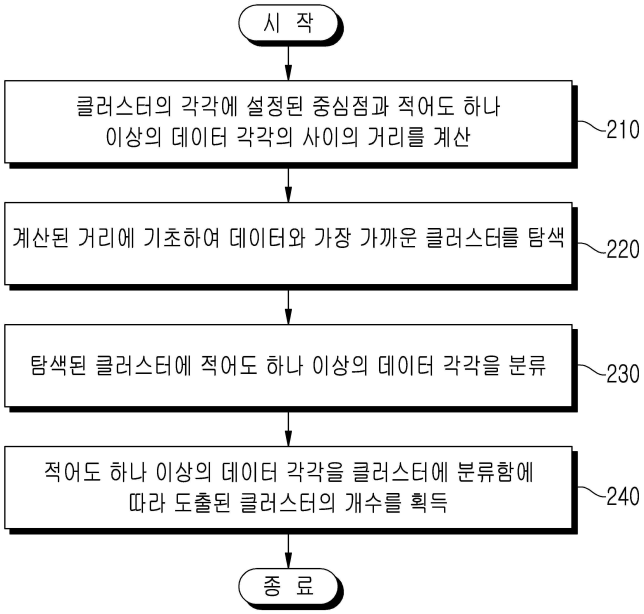
[0054] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

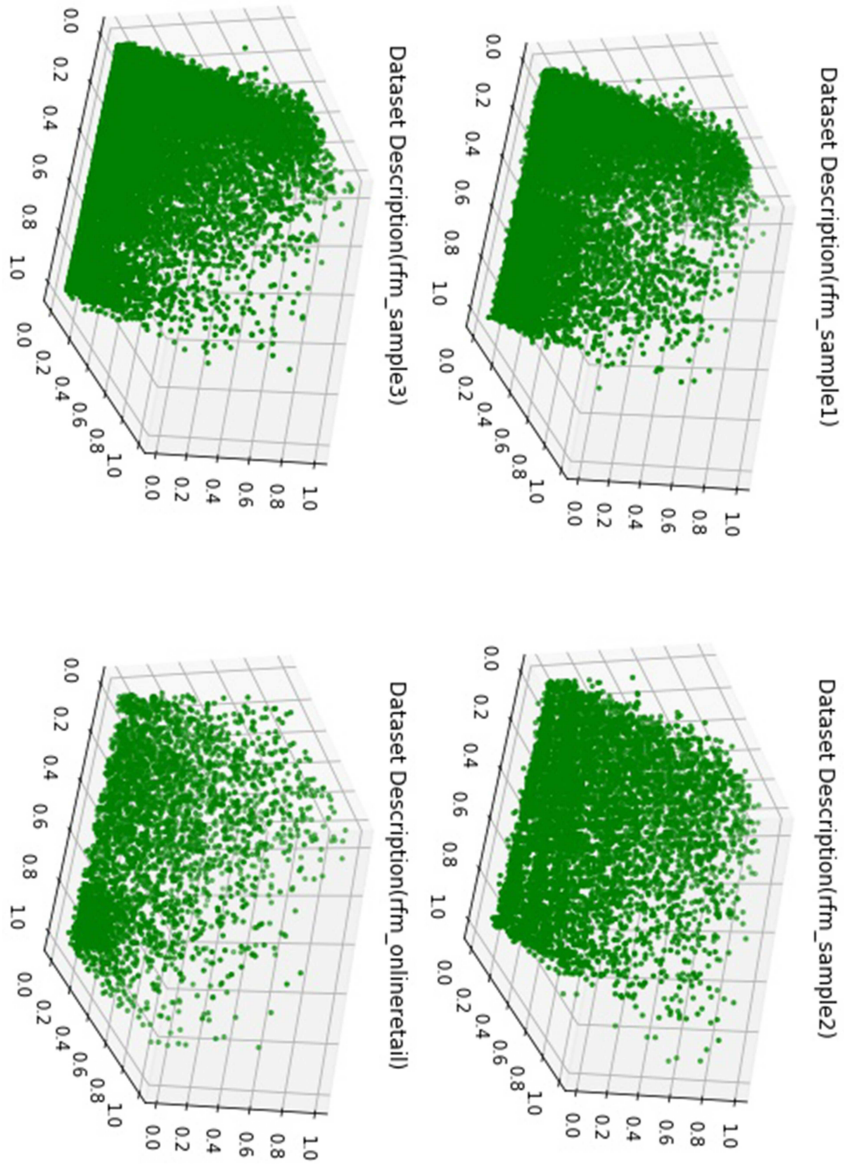
도면1



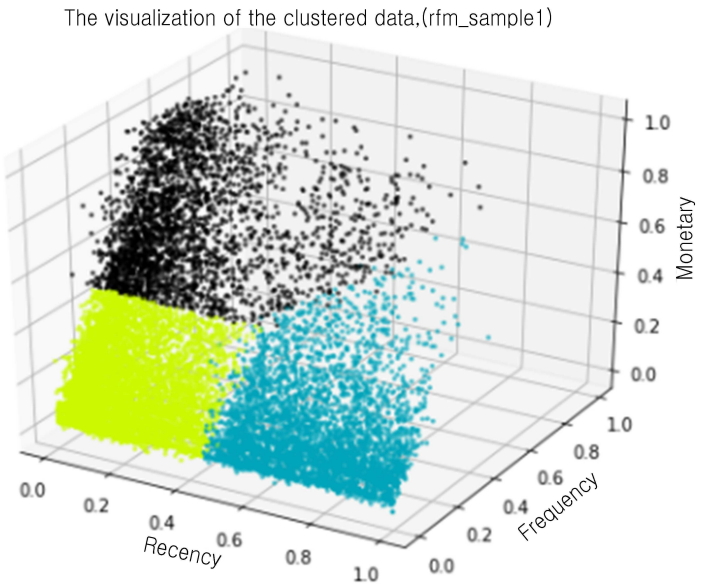
도면2



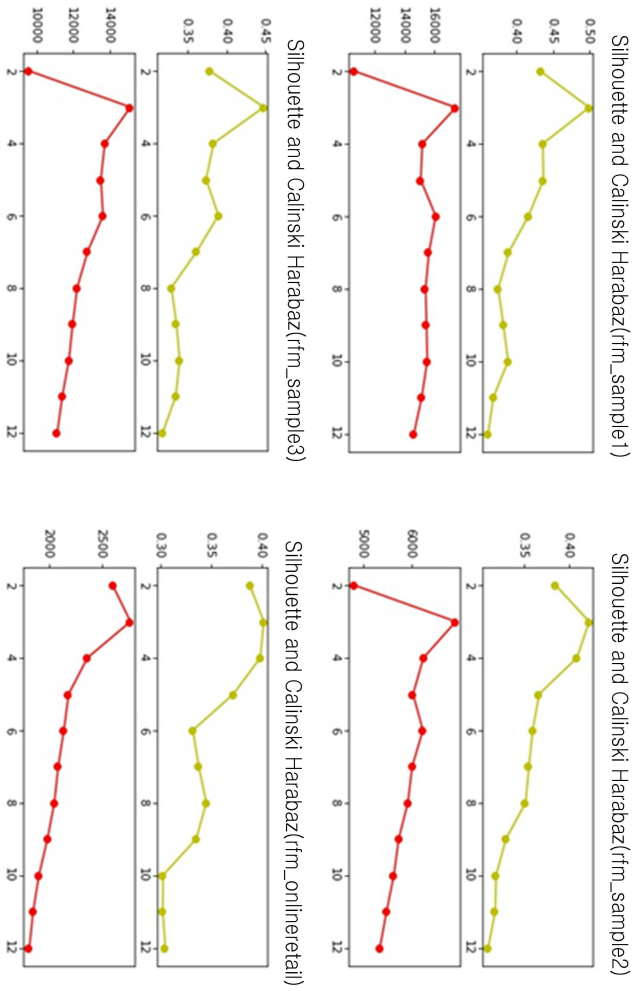
도면3


Dataset Description(rfm_sample1)
Dataset Description(rfm_sample2)
Dataset Description(rfm_sample3)
Dataset Description(rfm_onlineretail)

도면4


The visualization of the clustered data,(rfm_sample1)
Recency
Frequency
Monetary

도면5


Silhouette and Calinski Harabaz(rfm_sample1)
Silhouette and Calinski Harabaz(rfm_sample3)
Silhouette and Calinski Harabaz(rfm_sample2)
Silhouette and Calinski Harabaz(rfm_onlineretail)