# 1. 데이터마이닝이 적용된 사례를 찾아 소개해 보시오.

|  |
| --- |
| **(1) 미국 국세청, 탈세 방지 시스템 통한 국가 재정 강화 - 탈세 및 사기로 인한 국가의 재정 위기 가능성 증가**    [추진내용]  대용량 데이터와 다양한 기술을 결합하여 탈세 및 사기 범죄 예방 시스템 구축    사기방지 솔루션    소셜 네트워크 분석    데이터 통합 및 지능형 감시 시스템 구축    [효과]    세금 누락 및 불필요한 세금 환급 절감의 효과 발생    과학적 데이터를 근거로 탈세 조사를 진행함으로 탈세자 수 감소    향후 범죄 미 및 탈세 관련 사건 미연에 방지    1.png  **(2) 일본, 센서데이터를 활용한 지능형 교통안내 시스템 - 실시간으로 GPS 데이터를 분석하여 최적의 교통 정보를 사용자에게 전달하는 서비스**    [추진내용]  GPS로부터 자동차 주행의 스피드를 계산하여 교통 정보 수집. 지능형 교통 정보 시스템의 발전    택시 및 정보 제공에 동의한 내비게이터 사용자로부터 얻어진 교통 정보를 이용    수집된 교통 정보를 바탕으로 실시간 최적의 교통 안내 서비스 제공    [효과]    실시간 교통 정보 공유로 최적의 교통 안내 서비스 가능    교통 체증으로 인한 불필요한 에너지 낭비 방지 및 효율 증대    2.png |

# 2. 군집1 에 (0,0)와 (0,1)이 속하고, 군집2에 (1,0)과(2,1)이 속한다고 하자.

맨해튼 거리 측정법에 따른 거리 계산

|  |  |
| --- | --- |
|  | 군집1 (0,0), (0,1) |
| 군집2 (1,0), (2,1) | (0,0)와 (1,0)의 거리=> | 0 – 1 | + | 0 – 0 | = 1  (0,0)와 (2,1)의 거리=> | 0 – 2 | + | 0 – 1 | = 3  (0,1)와 (1,0)의 거리=> | 0 – 1 | + | 1 – 0 | = 2  (0,1)와 (2,1)의 거리=> | 0 – 2 | + | 1 – 1 | = 2 |

## (1)단일연결법

|  |
| --- |
| 최대연결방법이므로 군집간의 군집개체간의 최단 거리가 군집간 거리가 되므로  (0,0)와 (1,0)의 거리=> | 0 – 1 | + | 0 – 0 | = 1 |

## (2) 완전연결법

|  |
| --- |
| 최장연결법이므로 군집간 군집개체간의 최장 거리가 군집간 거리가 되므로  (0,0)와 (2,1)의 거리=> | 0 – 2 | + | 0 – 1 | = 3 |

## (3) 평균연결법을 각각 이용하여 군집1과 군집2의 거리를 구하고 결과를 비교해 보시오.

|  |
| --- |
| d( P, Q ) = ( (0,0)와 (1,0)의 거리 ) + ( (0,0)와 (2,1)의 거리 ) + ( (0,1)와 (1,0)의 거리 ) + ((0,1)와 (2,1)의 거리 ) / ( 군집1의 개수 + 군집2의 개수 )  = ( 1 + 3 + 2 + 2 ) / ( 2 + 2 ) = 2  군집간 거리가 단일연결법이 가장 짧고 완전연결법이 가장 길고, 평균연결법이 단일연결과 완전연결법의 중간값이 됩니다. |

# 3. 나무모형과 연관규칙의 장단점을 정리하고, 양 기법 간의 특징을 비교하시오.

|  |  |
| --- | --- |
|  | 장단점 |
| 나무모형 | 장점  - 쉽게 이해할 수 있는 모형  - 변수간의 교호작용 관계를 잘 나타냄  - 변수종류에 상관없이 사용 가능  - 계산속도가 빠르고 대형자료 처리에 용이  단점  - 변수간의 교호작용이 지나치게 강조  - 재귀적인 알고리즘을 사용하므로 초기분할에 큰 영항  - 이산형변수에 대하여 수준이 많을 경우 결과가 정확하지 않음  - 과대적합된 모형이 적성되기 쉬워서 예측력이 낮을 가능성이 높음 |
| 연관규칙 | 장점 : 사용하기 쉽고 결과가 명확, 이해하기 쉽음  단점 : 품목수가 증가함에 따라 계산량이 매우 증가 |

|  |  |
| --- | --- |
|  | 특징 정리 |
| 나무모형 | 지도학습으로 독립변수로 예측변수를 예측하는 분석  항상 목적변수에 대한 결과 구함  하나의 변수에 대한 결과를 구함  규칙을 동시에 구하여 고객을 겹치지 않는 여러 개의 집단으로 나눔  고객을 겹치지 않게 여러집단으로 나눔 |
| 연관규칙 | 자율학습으로 주로 상품구매와 같은 사건을 중심으로 분석  목적변수 없이 특성의 조합으로 규칙 표현  여러 변수 또는 사건에 대해여 구함  규칙들을 각기 구하는데 규칙들을 구할 때 서로 영향을 주지 않음  하나의 고객이 여러 개의 규칙에 해당될 수 있음 |

# 4. 1970년 미국의 도시인구 100,000명당 범죄비율에 대한 자료

## (1) 위의 자료에 대하여 단일연결법, 완전연결법, 평균연결법에 따른 계층적 군집화를 시행하고 각 방법에 대한 나무형 그림을 그려보시오.(유클리디안 거리 사용)

|  |
| --- |
| ex4 <- read.table( "D:/work\_R/Datamining\_work/ex4.txt", header=T )  crime\_rate <- as.matrix( ex4[ , -1] )  single <- hclust( dist(crime\_rate, method="euclidean"), method="single" )  complete <- hclust( dist(crime\_rate, method="euclidean"), method="complete" )  average <- hclust( dist(crime\_rate, method="euclidean"), method="average" )  par(mfrow=c(2,2))  plot(single, main = "1970년 미국의 범죄비율-단일연결법" )  groups <- cutree(single, k=5)  rect.hclust(single, k=5, border="red")  plot(complete, main = "1970년 미국의 범죄비율-완전연결법" )  groups <- cutree(complete, k=5)  rect.hclust(complete, k=5, border="red")  plot(average, main = "1970년 미국의 범죄비율-평균연결법" )  groups <- cutree(average, k=5)  rect.hclust(average, k=5, border="red") |

## (2) 분할분석을 이용한 계층적 군집화를 시행하시오. 나무형 그림을 산출하고 (1)에서의 결과와 비교하시오. (유클리디안 거리 사용)

|  |
| --- |
| install.packages("cluster")  library(cluster)  diana <- diana( crime\_rate , metric="euclidean")  par(mfrow=c(1,2))  plot(diana, main = "1970년 미국의 범죄비율-DIANA를 이용한 계층적 군집화" )  groups <- cutree(diana, k=5)  rect.hclust(diana, k=5, border="red")    **분할분석(DIANA) 군집은 단일연결법으로 군집화 결과와는 상이하고 완전연결과 평균연결법의 결과와 비슷함.** |

## (3) 위의 (1)과 (2) 항목을 유클리디안 거리 대신 맨해튼 거리를 사용하여 분석하고 차이점이 있는지 결과를 비교하시오.

|  |
| --- |
| **단일연결법에서는 유클리다인에서는 12번이 단일군집이 되지만, 맨해튼에서는 8번이 단일군집이됨**  **유클리디안 거리에서는 완전연결법과 평균연결법이 동일한 결과가 나오지만, 맨해튼 거리에서는 다른 결과가 나옴.**  **DIANA에서는 군집1에 2번과 12번이 거리에 따라서 포함되는 것이 달라짐** |

# 5. 다음은 어느 소매점의 거래자료이다. 이를 바탕으로 연관규칙 분석을 하고자 할 때, 물음에 답하시오.

## (1) 주어진 거래에 대하여 연관규칙 ‘맥주→콜라’에 대한 지지율, 신뢰도, 향상도를 구하시오.

|  |
| --- |
| 지지율 : A -> B의 지지율 = A와 B를 포함한 거래수 / 전체거래수  2 / 5 = 0.4 |
| 신뢰도 : A -> B의 신뢰도 = A와 B를 포함한 거래수 / A를 포함한 거래수  2 / 3 = 0.67 |
| 향상도 : A -> B의 신뢰도 / B를 포함한 거래비율  0.67 / 0.6 = 1.12 |

## (2) 주어진 거래에 대하여 연관규칙 ‘맥주&땅콩→치즈’에 대한 향상도를 구하고 의미를 밝히시오.

|  |
| --- |
| 신뢰도 : 1 / 1 = 1  향상도 : 1 / 0.8 = 1.25  맥주&땅콩를 구매한 고객은 반드시 치즈를 사고 이사실을 알고 버터에 대한 마케팅을 한다면, 이 사실을 모들 때보다 25% 증가 |