프로젝트 #3: DB Mining

본 프로젝트는 주어진 데이터를 이용하여 연관 분석 및 의사결정 나무 모델을 생성하고 결과를 분석하는 것을 목적으로 한다. 프로젝트 #2에서 진행했던 사이트 A의 데이터 및 추가 제공 데이터를 사용한다.

해당 프로그램은 python와 MySQL, Orange3, scikit-learn를 사용하여 구현하여야 하며, 다음 요구 조건을 만족하여야 한다. 본 프로젝트는 크게 두 부분으로 나뉘게 된다.

연관 분석 – (R1), (R2), (R3)

의사 결정나무 – (R4), (R5), (R6)

Part I.

(R1) 프로그램은 연관 분석을 위한 테이블 뷰 TagMatrix를 MySQL 상에 생성한다. 프로젝트 #2의 questionPosts.csv (질문 게시물)데이터를 사용해 만든 테이블의 이름을 'questionposts'라고 하자. questionposts에는 질문 게시물에 대한 tagging이 되어 있는 column이 포함되어 있다. 여기서, 추가적으로 주어진 데이터 tagname.csv를 사용하여 TagMatrix를 생성한다. (tagname.csv는 questionPosts.csv의 Tags column에 포함된 모든 Tag의 인용 횟수를 count하여, 내림차순으로 정렬한 파일이다) 예를 들면 Orders의 데이터가 아래와 같이 주어졌을 경우,

Id	PostId	AcceptedAnswerId	ViewCount	Title	Tags
1	1	6	1728	a	<r><dataset><pca><score>
2	2	29	8198	b	<r><machine-learning>
3	3	1	3613	c	<anova><machine-learning>

아래와 같은 TagMatrix View가 생성된다.

	Tags에 'r' 이 포함되어 있음 (0혹은 1)	Tags에 'machine-learning'이 포함되어 있음 (0혹은 1)	Tags에 'anova'가 포함되어 있음 (0혹은 1)
Id 1	1	0	0
Id 2	1	1	0
Id 3	0	1	1

(R2) 프로그램은 (R1)을 통해 얻은 TagMatrix 데이터로부터 Tag 간의 연관 분석을 실시하고 결과를 출력한다. 각각의 연관 분석은 다음의 조건을 만족해야 한다.

사용 알고리즘	FP-Growth
평가척도	Confidence
사용 Tag 개수	인용 수 상위 100개
Min_support	0.01
Min_confidence	0.05

(R3) 분석 모델의 결과를 정량적인 관점과 정성적인 관점에서 분석하시오. (정량적 분석에는 confidence 와 lift를 활용한 분석이 포함되어야 한다.)

Part II.

(R4) 사이트 A의 사용자는 고유의 평판 수치를 가지고 있다. 이 때, 다양한 feature를 바탕으로 평판 수치를 추론할 수 있는 의사결정나무를 그려보고자 한다. 이를 위해 먼저 다음과 같은 column이 포함되는 ReputStatMatrix 뷰를 생성한다.

- UserId (from userInfo.csv / Id)
- Reputation (from userInfo.csv / Reputation) → 값이 110보다 큰 row 만 뷰에 포함시킨다.
- NumOfPosts (from posts.csv) → 게시물 작성 수
- NumOfComments (from comments.csv) → 댓글 작성 수
- NumOfBadges (from badges.csv) → 획득한 뱃지 수

UserId	Reputation	NumOfPosts	NumOfComments	NumOfBadges
33	221	3	1	2
34	185	8	1	8
35	101	3	2	1

(생성되는 ReputStatMatrix의 일부)

(R5) (R4)에서 구현한 ReputStatMatrix 뷰 데이터로부터 Reputation이 180을 초과하는지 판단하는 의사결정나무를 생성하고자 한다. Node impurity 측정 방식을 ‘gini’와 ‘entropy’ 두 가지로 하여 각각 의사결정나무를 생성하며, 이 때 결과는 graphviz library를 통해 저장한다. 본 문제에서 만드는 의사결정나무의 속성을 정리하면 다음과 같다.

A. 사용 library : sklearn.tree.DecisionTreeClassifier
B. Binary Classifier 사용
C. Criterion : gini / entropy
D. min_samples_split =10
E. graphviz 출력 format = ‘png’
F. 나머지 속성은 Default값으로
G. 분석 목표: Reputation
   i. Reputation이 180 초과
   ii. Reputation이 180 이하

위와 같은 조건에서 만들어진 모델을 간단히 분석하시오. 또한, 다음과 같은 feature를 가지는 사용자는 위 의사결정나무 (‘gini’ 및 ‘entropy’) 모델에서 어떤 class로 분류되고, 그 때 각 class로 분류될 확률을 얼마인지 구하시오.

UserId	Reputation	NumOfPosts	NumOfComments	NumOfBadges
1000000	?	5	5	5
1000001	?	2	6	18
1000002	?	6	3	10

(R6) (R5)에서 만든 의사결정나무를 발전시키기 위해 input feature를 추가하고자 한다. 기존 NumOfPosts, NumOfComments, NumOfBadges에 다른 feature를 추가하여 뷰를 생성하고 (이 때 뷰의 이름은 ReputStatMatrix2로 하며, 추가하는 attribute(feature)의 수는 최대 2개로 한다), 이를 활용하여 의사결정나무를 생성한다. 의사결정나무의 속성은 (R5)와 동일하게 하며, 마찬가지로 graphviz library 를 통해 저장한다. 또한 생성된 모델을 분석하고 이를 (R5)의 모델과 간단히 비교하시오.

구현한 프로그램 전체 소스(주석 포함)와 보고서, 발표자료 및 graphivz 의

사결정나무 그림 파일을 함께 제출하여야 한다. 프로그램 소스는 두 개의 함수 ‘association’(: part1에 대한 solution), ‘decisiontree1’(: part2 (R4),(R5)에 대한 solution), ‘decisiontree2’(: part2 (R6)에 대한 solution)로 구성되어 있어야 하고, 이 때 각각의 함수는 host, user, password를 input parameter로 받아 이를 활용하여 MySQL에 연결하는 객체인 pymysql.connect 객체를 갖고 있어야 한다. 또한 각각의 함수에 사용할 csv를 읽어들여서 database에 저장하는 프로그램 또한 포함되어 있어야 하며, 이 때 table은 3NF 조건 및 domain constraint 조건을 지키지 않아도 된다. 제출할 보고서에는 소스 프로그램 설명, 프로그램 실행 결과, 모델의 결과 및 분석을 포함하여야 한다. 소스 제출의 경우 단순 python 코드만 제출해서는 안되며 프로젝트 파일 전체를 제출해야 한다.

본 프로젝트의 발표시간은 4분이며 보고서와 발표자료는 12월 14일 0시까지 ETL에 업로드 해야 한다. **보고서는 최대한 간략하고 명료하게 작성한다.** 또한 발표날 수업시간에 보고서의 하드 카피를 담당 강의자에게 제출해야 한다. (하드카피의 내용과 ETL에 제출된 내용은 정확히 같아야 한다.)

채점 기준:

- DB mining 프로그램의 requirements 만족 여부: 90%
- 보고서 품질: 10%