T.R.
GEBZE TECHNICAL UNIVERSITY
FACULTY of ENGINEERING
DEPARTMENT of COMPUTER ENGINEERING

DBSCAN ALGORTIHM IMPLEMENTATION

CSE 454 DATA MINING
ASSIGNMENT 1 REPORT

STUDENT
Şeyda Nur DEMİR
12 10 44 042

LECTURER
Burcu YILMAZ

TEACHING ASSISTANT
-

KOCAELİ, 2020

1.DESCRIPTION

1.1.Requirements :

1. Implement DBSCAN model. You must use the algorithm mentioned in the book. You can use any programming language. Find a dataset to present the results. (You can not use any code from anywhere.)
2. Prepare an assignment report showing extracted clusters for at least 3 values of each parameter.
3. In the report, write a discussion about how the parameters effect the results.
4. In the report, give a technique to automatically decide on the parameters of DBSCAN?

1.2.Deadline :

- 06.12.2020 23:55
- You must upload your assignment to moodle. I will add the assignment to moodle.

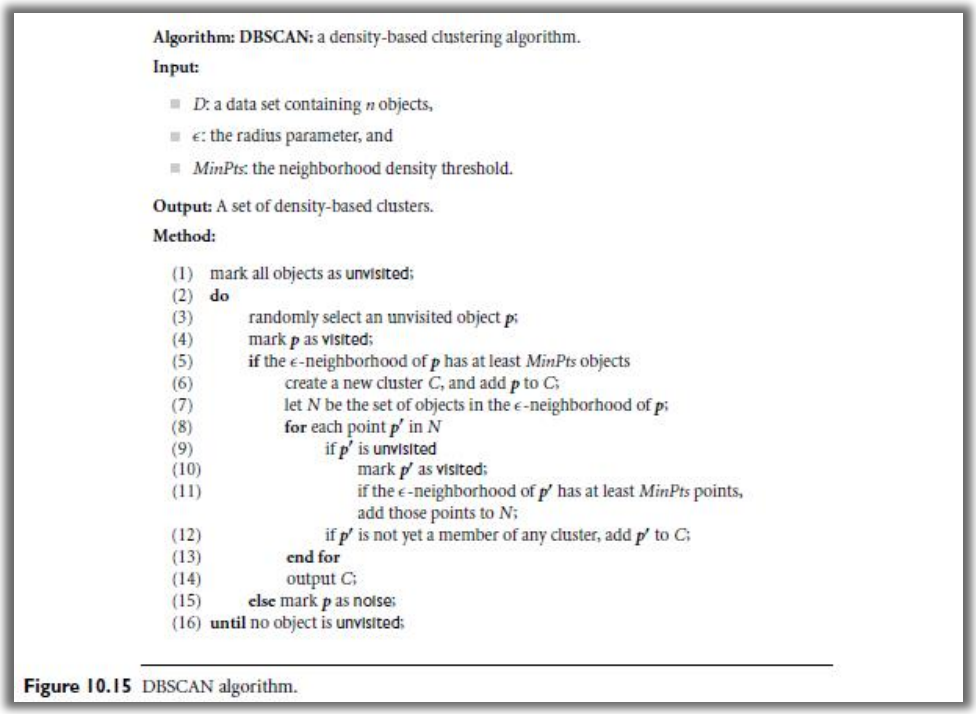
1.3.Demo :

- Before 15.01.2021 (Please ask an appointment from me to attend the demo)
- You must present your code in the demo section. If you do not attend the demo section, you will get zero from the assignment, even if you upload your assignment.

2.DBSCAN ALGORITHM

I implemented DBSCAN algorithm according to the method mentioned in the our course book (at pages 471,472, and 473) in my home work.

DBSCAN algorithm method is shown as below :



Algorithm: DBSCAN: a density-based clustering algorithm.

Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

```
(1)  mark all objects as unvisited;
(2)  do
(3)      randomly select an unvisited object p;
(4)      mark p as visited;
(5)      if the ε-neighborhood of p has at least MinPts objects
(6)          create a new cluster C, and add p to C;
(7)          let N be the set of objects in the ε-neighborhood of p;
(8)          for each point p' in N
(9)              if p' is unvisited
(10)                 mark p' as visited;
(11)                 if the ε-neighborhood of p' has at least MinPts points,
                     add those points to N;
(12)             if p' is not yet a member of any cluster, add p' to C;
(13)         end for
(14)         output C;
(15)     else mark p as noise;
(16) until no object is unvisited;
```

Figure 10.15 DBSCAN algorithm.

Source : https://www.elsevier.com/books/data-mining-concepts-and-techniques/han/978-0-12-381479-1

Pseudocode of this method is shown as below :

```
DBSCAN (SetOfPoints, Eps, MinPts)

// SetOfPoints is UNCLASSIFIED
  ClusterId := nextId(NOISE);
  FOR i FROM 1 TO SetOfPoints.size DO
    Point := SetOfPoints.get(i);
    IF Point.ClId = UNCLASSIFIED THEN
      IF ExpandCluster(SetOfPoints, Point,
            ClusterId, Eps, MinPts) THEN
        ClusterId := nextId(ClusterId)
      END IF
    END IF
  END FOR
END; // DBSCAN
```

```
ExpandCluster(SetOfPoints, Point, ClId, Eps,
              MinPts) : Boolean;
  seeds:=SetOfPoints.regionQuery(Point,Eps);
  IF seeds.size<MinPts THEN // no core point
    SetOfPoint.changeClId(Point,NOISE);
    RETURN False;
  ELSE   // all points in seeds are density-
         // reachable from Point
    SetOfPoints.changeClIds(seeds,ClId);
    seeds.delete(Point);
    WHILE seeds <> Empty DO
      currentP := seeds.first();
      result := SetOfPoints.regionQuery(currentP,
                                        Eps);
      IF result.size >= MinPts THEN
        FOR i FROM 1 TO result.size DO
          resultP := result.get(i);
          IF resultP.ClId
              IN (UNCLASSIFIED, NOISE) THEN
            IF resultP.ClId = UNCLASSIFIED THEN
              seeds.append(resultP);
            END IF;
            SetOfPoints.changeClId(resultP,ClId);
          END IF; // UNCLASSIFIED or NOISE
        END FOR;
      END IF; // result.size >= MinPts
      seeds.delete(currentP);
    END WHILE; // seeds <> Empty
    RETURN True;
  END IF
END; // ExpandCluster
```

Source : https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf

Dataset that I used : COVID-19 Dataset (Confirmed and Deaths Columns)

Source : https://www.kaggle.com/imdevskp/corona-virus-report

Parameters I tried in my homework			
ε	0.3	1.0	5.0
minPts	3	5	10

Discussion about how the parameters effect the results :

I see …

A technique to automatically decide on the parameters of DBSCAN :

The technique ..

END OF THE REPORT

UPDATED : 06.12.2020 03:50

STUDENT
Şeyda Nur DEMİR
12 10 44 042

LECTURER
Burcu YILMAZ

TEACHING ASSISTANT
-

KOCAELİ, 2020