Birla Institute of Technology and Science-Pilani, Hyderabad Campus

SEMESTER-II 2018-2019

Data Mining (CS F415)

DBSCAN and Location Outlier Factor

BY

Garvit Jain 2016A7PS0080H
Aman Mehta 2016A7PS0066H
Ayushi Srivastava 2016A1PS0587H

UNDER THE SUPERVISION OF

Mrs. Aruna Malapati

Dataset

Credit Card fraud dataset

The Database contains different factor which may indicate fraud.

- Number of instances: 284806
- Number of attributes: 30

Pre-processing done on data

The data given in csv file is converted into a pandas dataframe. The pandas dataframe is then normalised and a dictionary of distances is formed. Both these data structures are dumped in pickle files for further use.

Formulas Used

- Euclidean distance :

$$d(p,q) = d(q,p)$$
$$= \sqrt{(q_1 - p_1)^2 + (q_3 - p_3)^2 + \cdots + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- DBSCAN :

1. Choose values for Minpts > 0 and Eps > 0

2. $A_i = \{x \in S : d(x_i, x) <= Eps\}$; i=1,2…,n

3. If $|A_i| <$ Minpts ignore the point

4. Take union of A_i and A_j if $A_i \cap A_j \neq \Phi$

5. Repeat 4 till no union take place

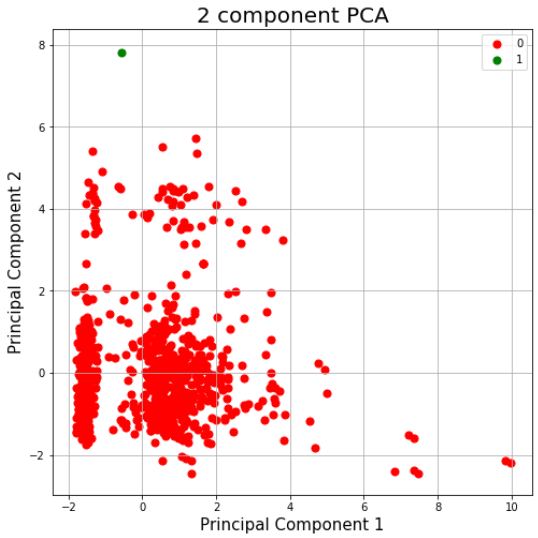
- LOCATION OUTLIER FACTOR :

$$\text{reachability-distance}_k(A, B) = \max\{\text{k-distance}(B), d(A, B)\}$$

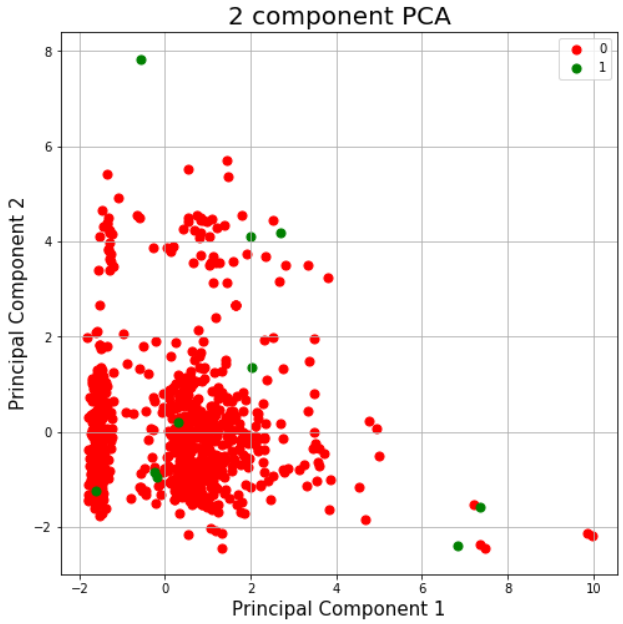
$$\text{lrd}_k(A) := 1/\left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|}\right)$$

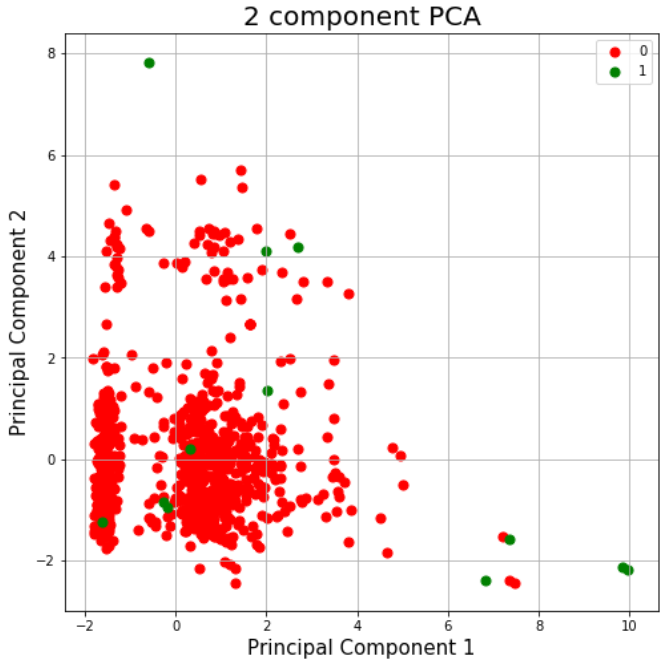
$$\text{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}(B)}{|N_k(A)|} / \text{lrd}(A)$$

Results:



Dbscan





LOCAL OUTLIER FACTOR