OUTLIER ANALYSIS

Submitted By :-

Prayank Mathur Nishant Raj
2012JE0719 2012JE0847
III Semester
Dept. of Computer Science & Engg.

What is an Outlier ?

Data objects which are grossly different from or inconsistent with the remaining set of data are called outliers.

Applications of Outliers

- Fraud Detection
- Unusual Usage of Credit Cards
- Medical Analysis
- Customized Marketing

Outlier Mining Problem

- It can be viewed as two Subproblems.

(1) Define what data can be considered as inconsistent.

(2) Find an efficient method to mine the outliers.

Methods of Outlier Detection

- Statistical Distribution-Based Outlier Detection
- Distance-Based Outlier Detection
- Density-Based Local Outlier Detection
- Deviation Based Outlier Detection

Distance Based Outlier Detection

- It was introduced to counter the main limitations imposed by statistical methods
- An object,**O,** in a dataset is a **distance based (DB) outlier** with parsameters *pct* and *dmin*, that is, a ***DB(pct,dmin)-outlier***.

Algorithms

- Index-based algorithm
- Nested-loop algorithm
- Cell-based algorithm
- ***N DoT***

N DoT

- We introduce a term *Nearest Neighbour Factor* (**NNF)** measure the degree of outlierness of a point.
- If *Nearest Neighbor Factor* of the point w.r.t majority of its neighbor is more than a threshold then the point is declared as a potential outlier.

Basic Terminologies

- K Nearest Neighbor (knn) Set
- Average knn distance
- Nearest Neighbor Factor

K Nearest Neighbor (knn) Set

- Let *D* be a dataset of and *x* be a point in *D*.
- For a natural number k and a distance function d, a set $Nnk(x)= \{q1 \epsilon D|d(x,q1) < d(x,q2), q2 \epsilon D\}$ is called knn of x if the following two conditions hold.

  (1) |Nnk| > k if q2 is not unique in D or |Nnk|=k otherwise.

  (2) |Nnk \ Nq2|=k-1, where Nq2 is the set of all q2 point(s).

Average knn Distance

- Let NNk be the knn of a point x in dataset D. Average knn distance of is the average of distances between x and q belongs to Nnk ,i.e.,
- Average knn distance(x)= ∑q d(x,q| qϵNnk)/|NNk|
- Average knn distance of a point x is the average of distances between x and its knn.
- If Average knn distance of x is less to compared to other point y, it indicates that *x*'s neighborhood is more densed compared to that of y.

Nearest Neighbor Factor(N N F)

- Let x be a point in D and Nnk(x) be the knn of x .
- The N N F of x with respect to qєNNk(x) is the ratio of d(x,q) and Average knn distance of q.
- NNF(x,q)=d(x,q)/Average knn distance.

How it Works ?

- Given a dataset D, it calculates knn and Average knn distance for all points in D.
- In the next step, it computes *Nearest Neighbor Factor* for all points in the dataset using the previously calculated *knn* and *Average knn Distance.*
- *NDoT* decides whether x is an outlier or not based on a voting mechanism.

- Votes are counted based on the generated NNF values with respect to all its k nearest neighbors.

- If NNF(x,q | q ϵ Nnk(x)) is more than a threshold value(=1.5 in most experiments), x is considered as an outlier with respect to q.

- Subsequently, a vote is counted for x being an outlier point. If the number of votes are at least 2/3 of the number of nearest neighbors then x is declared as an outlier point.