OUTLIER

L'analisi degli outlier non è da confondere con l'analisi del rumore.
Il rumore è dato da errori random, e dovrebbero essere eliminati perchè creano disturbo.
Gli outlier, d'altro canto, sono dei punti che violano il trend dei normali dati e possono avere dei significati intrinsechi non banali.

GLOBAL OUTLIER

Sono dati da quelli oggetti che si distaccano significativamente dal resto del DB.

CONTEXTUAL OUTLIER

Oggetti che deviano significativamente rispetto a un contesto (Temperatura di 35° in inverno ha un significato diverso rispetto all'estate).
Gli attributi del DB vengono suddivisi in:

1) Contextual attributes: definiscono il contesto [time, location]
2) Behaviour attribute: caratteristiche dell'oggetto [temperat.]

COLLECTIVE OUTLIER

Sono un gruppo di oggetti che non seguono lo stesso trend degli altri punti del DB. Hanno la particolarità che presi da soli magari non sono considerati outlier, ma in gruppo sì.

CHALLENGES

1) Come modellare gli oggetti normali e gli outlier poiché è difficile riconoscere tutti i tipi di outlier
2) Ogni applicazione ha i propri outlier
3) Distinguere il rumore dagli outlier
4) Capire come mai alcuni punti sono outlier.

METODI DI RILEVAMENTO DEGLI OUTLIER

1) SUPERVISIONATO

Anche quando i dati presentano una classe etichetta si può pensare di addestrare un modello che riconosca gli outlier, o viceversa uno che riconosca i dati normali.

2) NON SUPERVISIONATO

Qui bisogna esaminare il DB assumendo che i dati normali possano essere clusterizzati in gruppi multipli ognuno corrispondente a determinate caratteristiche comuni. In questo contesto gli outlier si suppone siano lontani da questi gruppi di punti normali.

Un difetto di questo metodo è la difficoltà di rilevare outlier collettivi, poiché possono condividere grandi correlazioni come i gruppi di dati normali.

3) SEMI - SUPERVISIONATO

Quando i dati etichettati sono pochi rispetto al totale, possiamo usare i dati etichettati e i dati non etichettati ma prossimi a quelli etichettati per allenare un modello che riconosca i dati normali. Gli outlier sono quelli che non si adattano al modello

4) METODO STATISTICO

Modellare statisticamente l'andamento dei dati normali Questi modelli dipendono fortemente da quando le assunzioni per modellare la statistica valgono nel mondo reale.

5) PROXIMITY - BASED

Semplicemente gli outlier hanno il vicinato più prossimo molto lontano. Quindi la prossimità di un outlier devia significativamente rispetto a quella degli altri dati.

6) CLUSTER-BASED

Si basano sul fatto che i dati normali appartengono a cluster densi e grandi, gli outlier invece a cluster piccoli e sparsi. Ci sono molti metodi basati sui metodi di cluster.