CS 512 Data Mining Principles

Outlier Detection

Hanghang Tong, Computer Science, Univ. Illinois at Urbana-Champaign, 2021

Suggested studying time: 3/19/2021-3/25/2021

Chapter 12. Outlier Detection

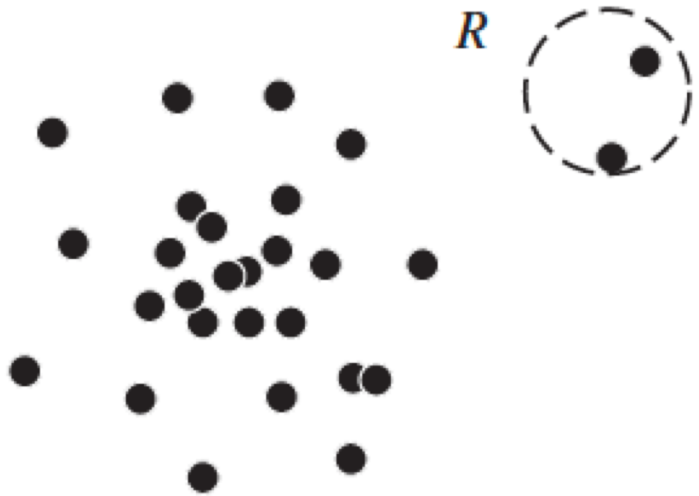
- Basic Concepts 
- Statistical Approaches
- Proximity-Based Approaches
- Reconstruction-Based Approaches
- Outlier Detection in High-Dimensional Data
- Summary

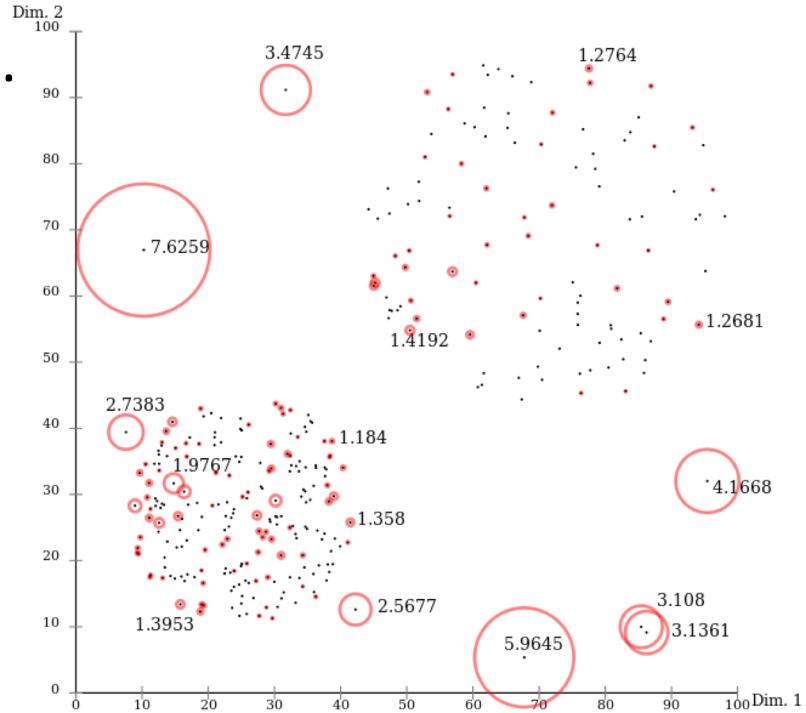
Basic Concepts

- What is Outliers?
- Types of Outliers
- Challenges of Outlier Detection
- Applications of Outlier Detection
- An Overview of Outlier Detection Methods

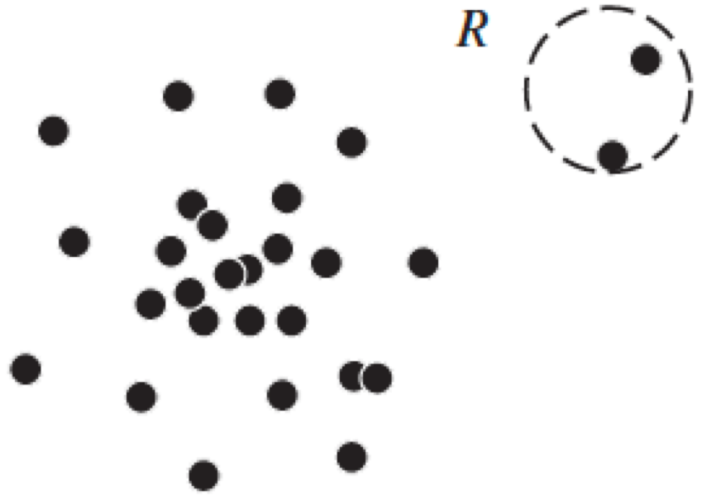
What is Outliers?

- Informally: things that are different from “normal” or “expected”
  - Abnormal vs. normal
- Formally: (Hawkins’ Definition of Outlier, 1980)
- “An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.”
- Many names: outlier, anomaly, outbreak, event, fraud, …

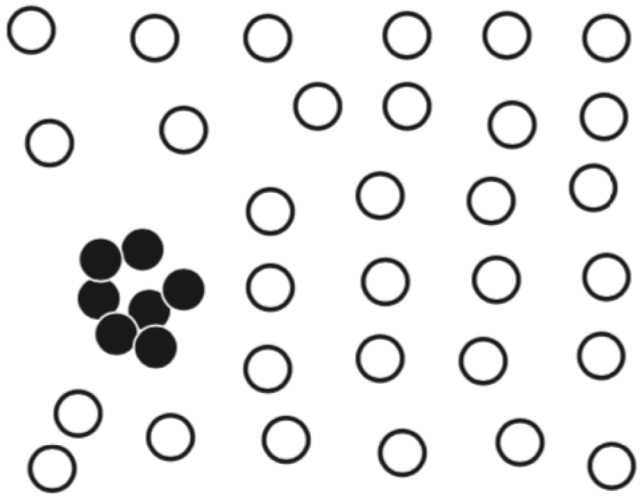




Types of Outliers



- Global Outliers (=point anomalies)
- Contextual Outliers (=conditional outliers)
  - Contextual attributes (e.g., date, location)
  - Behavioral attributes (e.g., temperature, humidity, and pressure)
  - Example:
    - 28C is an outlier for a Toronto winter.
    - 28C is not an outlier for a Toronto summer
- Collective Outliers (=group anomaly)
  - A group of data objects, as a whole, looks abnormal
  - But individual data object looks normal



Challenges of Outlier Detection

- Modeling normal objects and outliers
  - Difficulty in modeling normality, ambiguity between normal and abnormal
- Application-specific outlier detection
  - General-purposed techniques
- Noise vs. outliers
  - Noise: unavoidable, less interesting to the users, but make outlier detection more challenge (e.g., hide the outlier, blur the boundary, mislead detection)
- Interpretability
  - Why does the algorithm ‘think’ an object looks suspicious?

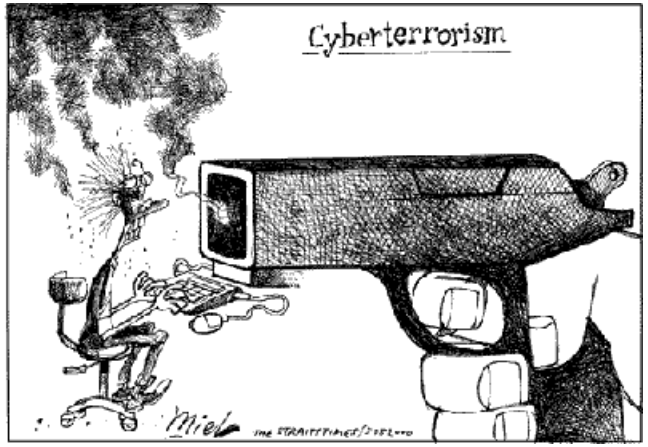
Outlier detection: Applications

Tax evasion

Credit card fraud

Healthcare fraud

Network intrusion



Outlier detection: Applications

Investment fraud

Click fraud

Spyware

Insurance fraud

Malicious cargo

Malware

Auction fraud

Damage detection

Fake reviews

Medical diagnosis

Email spam

False advertising

Performance monitoring

and many more…

Web spam

Insider threat

Image/video surveillance

An Overview of Outlier Detection Methods

- Taxonomy #1
  - Supervised Methods
  - Semi-Supervised Methods
  - Unsupervised Methods
- Taxonomy #2

  - Statistical Methods
  - Proximity-Based Methods
  - Reconstruction-Based Methods

Chapter 12. Outlier Detection

- Basic Concepts
- Statistical Approaches 
- Proximity-Based Approaches
- Reconstruction-Based Approaches
- Outlier Detection in High-Dimensional Data
- Summary

Statistical Approaches

- Basic Idea
  - Assume normal data are generated by a stochastic process
  - Data objects in low density regions are flagged as outlier
- Parametric Methods
  - the normal data objects are generated by a parametric distribution with a finite number of parameters
- Non-Parametric Methods
  - Does not assume a priori statistical model with a finite number of parameters

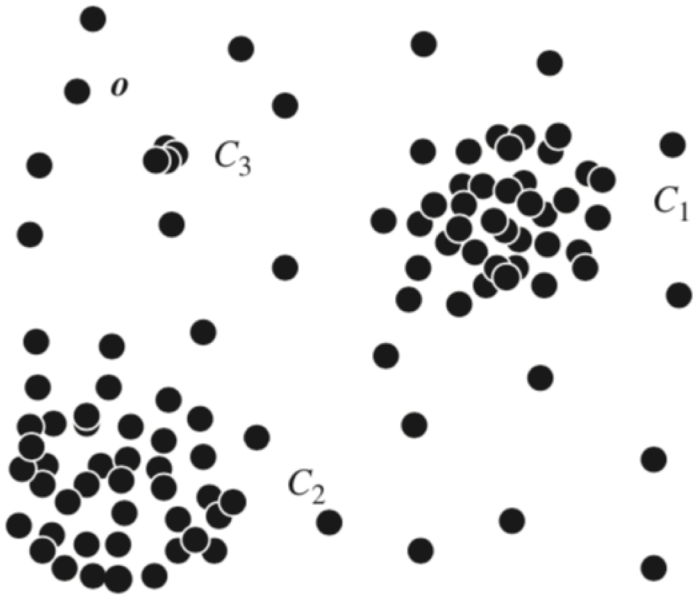
Parametric Methods

- Single Variable Data: Grubb's test (i.e., maximum normed residual test)
  - Normal data: Gaussian distribution $z = \frac{|x - \mu|}{\sigma}$
- Multi-variable Data
  - #1: Mahalanobis distance $MDist(\boldsymbol{o}, \bar{\boldsymbol{o}}) = (\boldsymbol{o} - \bar{\boldsymbol{o}})^T S^{-1} (\boldsymbol{o} - \bar{\boldsymbol{o}})$
  - #2: χ^2 – statistics

$$\chi^2 = \sum_{i=1}^{n} \frac{(o_i - E_i)^2}{E_i}$$

  - #3: mixture models

$$Pr(\boldsymbol{o}|\Theta_1, \Theta_2) = w_1 f_{\Theta_1}(\boldsymbol{o}) + w_2 f_{\Theta_2}(\boldsymbol{o})$$



Non-Parametric Methods

- Basic Idea:
  - Does not assume a priori statistical model with a finite number of parameters
- #1: Outlier Detection by Histogram
  - Construct histogram → data objects outside bins are outliers
  - Challenge: choose the right size of bin
- #2: Outlier Detection by Kernel Density Estimation
  - Kernel function: influence of a sample within its neighbor
    - $\int_{-\infty}^{+\infty} K(u)du = 1.$
    - $K(-u) = K(u)$ for all values of u.
  - Example: Gaussian kernel density estimation

$$K\left(\frac{x - x_i}{h}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}} \qquad \hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Chapter 12. Outlier Detection

- Basic Concepts
- Statistical Approaches
- Proximity-Based Approaches 
- Reconstruction-Based Approaches
- Outlier Detection in High-Dimensional Data
- Summary

Proximity-Based Approaches

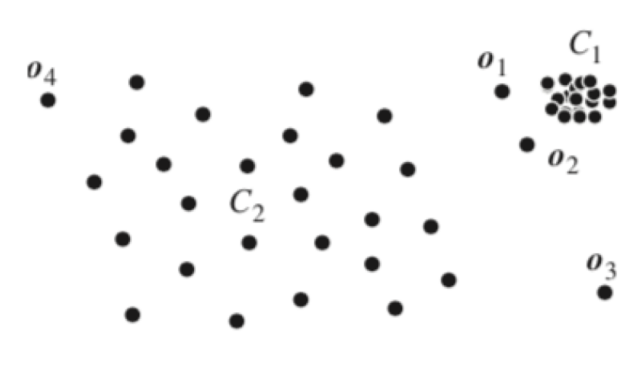
- Basic Idea
  - Intuition: objects that are far from others can be regarded as outliers
  - Assumption: the proximity of an outlier object to its nearest neighbors significantly deviates from the proximity of most other objects to their nearest neighbors
- Distance-Based Outlier Detection
  - Consult the neighborhood of a sample
  - Outlier: if there are not enough objects in its neighborhood
- Density-Based Outlier Detection
  - Compare the density of a sample with that of its neighbors
  - Outlier: if its density is relatively much lower than that of its neighbors.

Distance-Based Outlier Detection

- Basic Idea
  - Consult the neighborhood of a sample
  - Outlier: if there are not enough objects in its neighborhood
- Details
  - *r*: distance threshold; π: fraction threshold
  - *o* is a DB(r, π)-outlier if $\frac{\|\{o' | dist(o, o') \leq r\}\|}{\|D\|} \leq \pi$
  - Equivalent criteria: if $dist(o, o_k) > r$
    - o_k is the *k*-nearest neighbor of *o*
    - $k = \lceil \pi \|D\| \rceil$

Density-Based Outlier Detection

- Basic Idea
  - Compare the density of a sample with that of its neighbors
  - Outlier: if its density is relatively much lower than that of its neighbors.
- Details
  - *dist$_k$(o):* the distance between o and its k-nearest neighbor
  - *k*-distance neighborhood $N_k(o) = \{o' | o' \in D, dist(o, o') \leq dist_k(o)\}$
  - reachability distance $reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o, o')\}$
  - local reachability density $lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)}$
  - Local outlier factor $LOF_k(o) = \frac{\sum_{o' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} = \sum_{o' \in N_k(o)} lrd_k(o') \cdot \sum_{o' \in N_k(o)} reachdist_k(o' \leftarrow o)$



Chapter 12. Outlier Detection

- Basic Concepts
- Statistical Approaches
- Proximity-Based Approaches
- Reconstruction-Based Approaches 
- Outlier Detection in High-Dimensional Data
- Summary

Reconstruction-Based Approaches

- Basic Idea
  - Seek for an alternative, more succinct representation
  - Normal data: can be well reconstructed the original representation based on succinct representation
  - Outlier: cannot be well reconstructed based on the succinct representation
- Matrix Factorization Based Methods (for numerical data)
  - Succinct representation: matrix low-rank approximation
  - Goodness of reconstruction: reconstruction error
- Pattern-based Compression Methods (for categorical data)
  - Succinct representation: code table
  - Goodness of reconstruction: encoding length

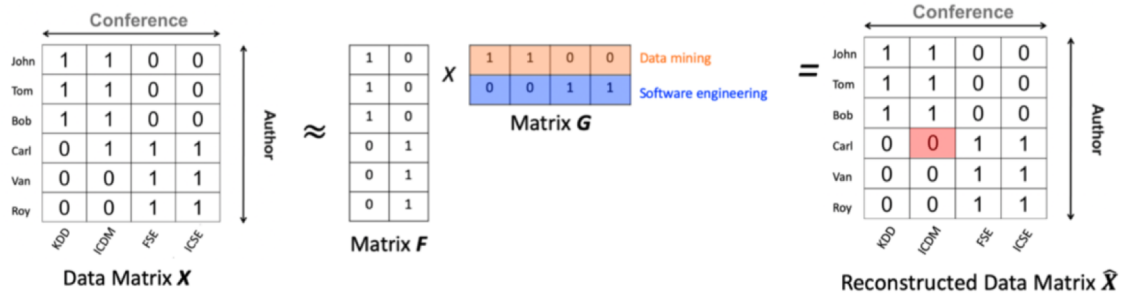
Matrix Factorization Based Methods

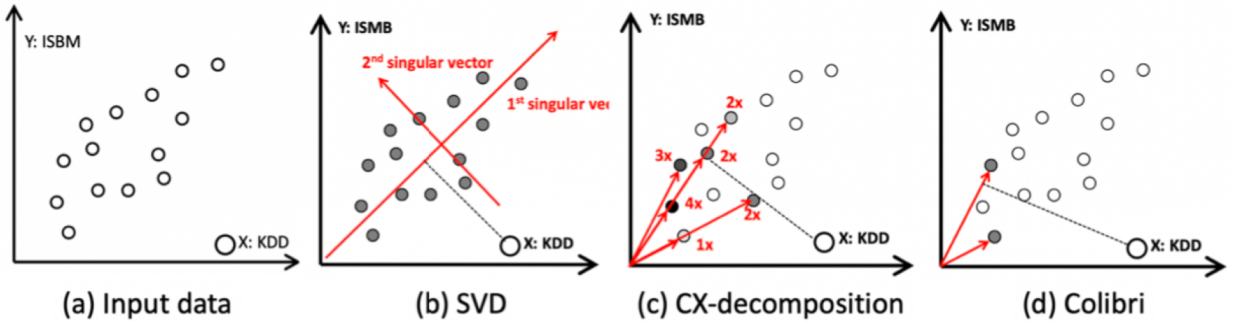
- Basic Idea
  - Succinct representation: matrix low-rank approximation
  - Goodness of reconstruction: reconstruction error
- Details
  - Step 1: represent the input data as a data matrix *X*
  - Step 2: approximate *X* by the multiplication of two low rank matrices *F* and *G*
  - Step 3: compute outlier-ness based on reconstruction error

$$r_i = \|X_i - \hat{X}_i\|^2 = \|X_i - \sum_{j=1}^{r} \boldsymbol{F}(i,j)\boldsymbol{G}(j,:)\|^2$$

  - Many choices for factorization
    - SVD, example-based methods,
    - NMF, NrMF, robust PCA and many more!

A numerical example





Pattern-based Compression Methods

- Basic Idea
  - Succinct representation: code table
  - Goodness of reconstruction: encoding length
- Details
  - Step 1: build the code table of the input data
  - Step 2: calculate the encoding length of input data
    - Longer encoding length → more outlying
- Remarks
  - Underlying math problem: MDL
  - finding optimal code table is NP-hard
  - Seek for effective heuristics

	Income	Credit	Purchase
John	High	High	High
Amy	High	High	High
Carl	Low	Low	Low
Mary	Low	Low	Low
Tom	High	Low	Medium
Jim	High	Low	Low

A numerical example

	I/H	I/M	I/L	C/H	C/M	C/L	P/H	P/M	P/L
John	1	0	0	1	0	0	1	0	0
Amy	1	0	0	1	0	0	1	0	0
Carl	0	0	1	0	0	1	0	0	1
Mary	0	0	1	0	0	1	0	0	1
Tom	1	0	0	0	0	1	0	1	0
Jim	1	0	0	0	0	1	0	0	1

Code word	Code	Usage	Code Length
[I/H, C/H, P/H]	01	2	2
[I/L, C/L, P/L]	10	2	2
[I/H, C/L]	11	2	2
[P/M]	001	1	3
[P/L]	010	1	3

Chapter 12. Outlier Detection

- Basic Concepts
- Statistical Approaches
- Proximity-Based Approaches
- Reconstruction-Based Approaches
- Outlier Detection in High-Dimensional Data
- Summary

Outlier Detection in High-Dimensional Data

- Main Challenges
  - Interpretation
  - Data sparsity
  - Data subspace
  - Scalability
- Overview of High-dimensional Outlier Detection
  - Finding Outliers in Subspaces
  - Outlier Detection Ensemble
  - Taming High-Dimensionality by Deep Learning

Finding Outliers in Subspaces

- Basic Idea
  - Find the subspace where certain data objects are flagged as outliers
  - Additional benefit: interpreting why and to what extent the object is an outlier
- Grid-based Subspace Outlier Detection Method
  - Intuition: consider projections of the data onto various subspaces. If the density of an area is lower than average, consider it as outliers
  - Details
    - Partition each dimension in ϕ equal-depth ranges, containing f faction objects (f=1/ ϕ)
    - Sparsity coefficient of k-dimension cube C (smaller coefficient indicates outlier-ness)
      - n: total # of objects
      - $n(C)$: # of objects in cube C
      - k: dimensionality of subspace

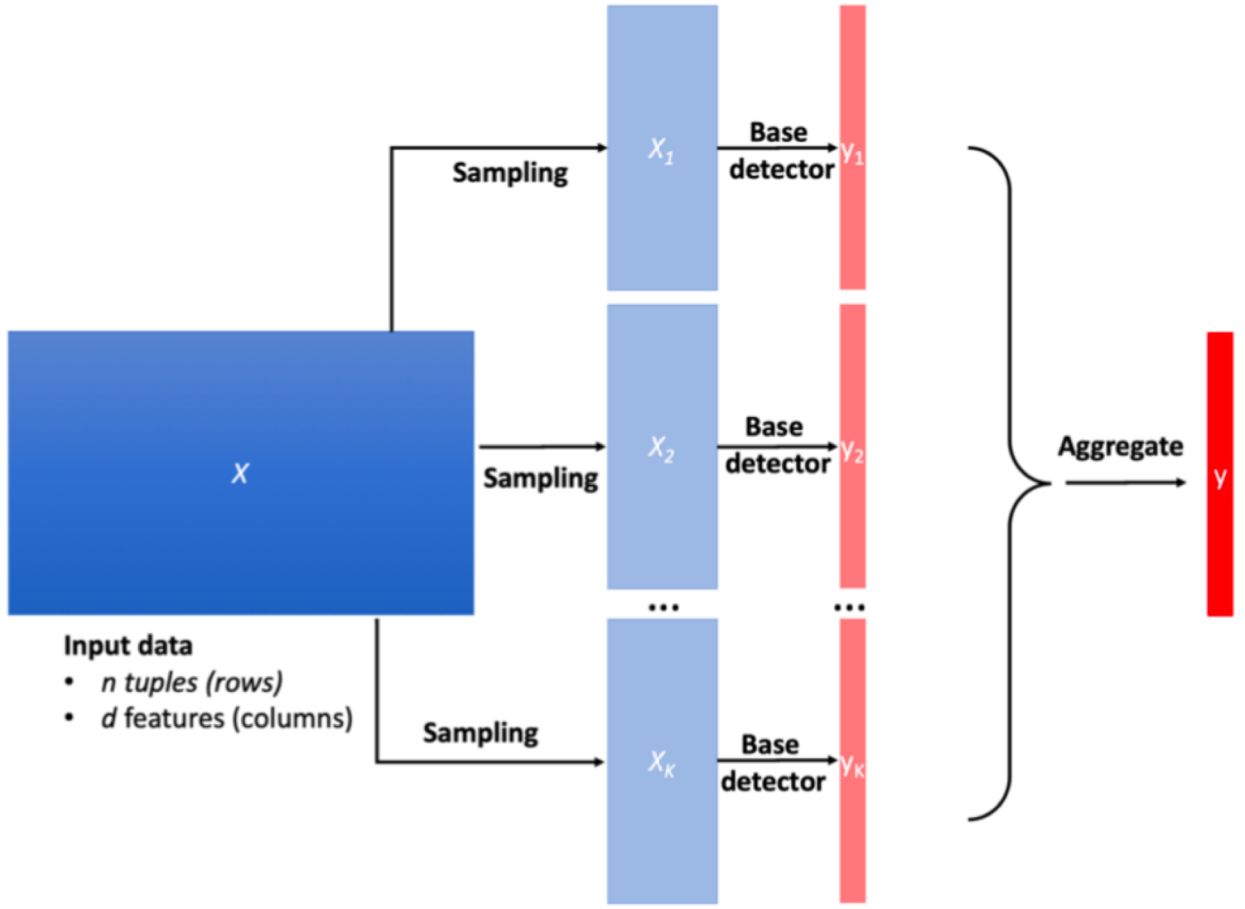
# of objects in C — Expected # of objects

$$S(C) = \frac{n(C) - f^k n}{\sqrt{f^k(1-f^k)n}}$$

Standard deviation

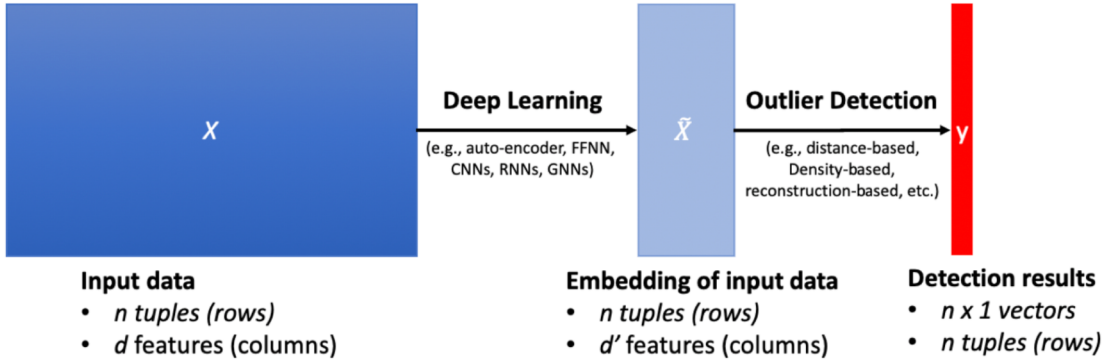
Outlier Detection Ensemble

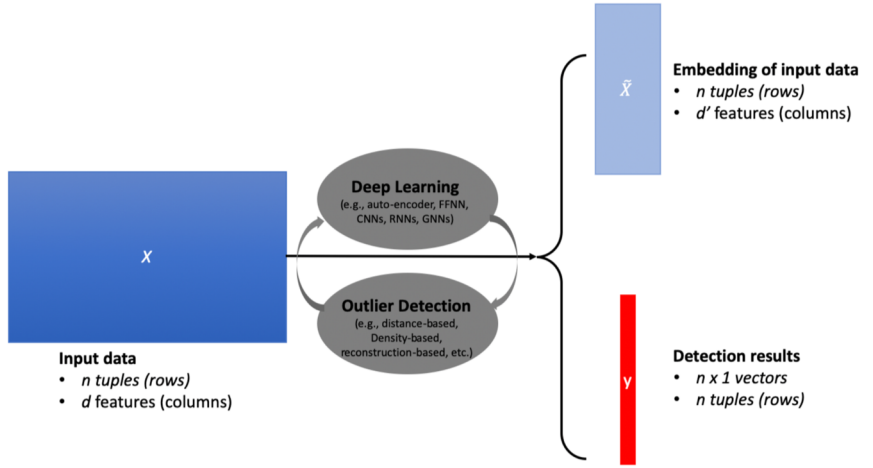
- Basic Idea
  - Each base detector assigns an outlier-ness score in a subspace
  - Aggregate the base detector results
- Find a Random Subspace
  - Feature bagging
  - Rated bagging
- Aggregate Base Detectors Results
  - Mean, max, etc.
  - Normalization is the key
    - Min-max, z-score, etc.



Taming High-Dimensionality by Deep Learning

- Why Deep Learning for high-D outlier detection?
  - Generate embedding of mush smaller dimensions
  - Capture the complex interaction between features
- Strategy #1: Deep learning as pre-processing
  - Output of the encoder of autoencoder
  - Proximity-based outlier detection
- Strategy #2: Integrate Deep Learning in outlier detection
  - Finding embedding that are tailored for outlier detection
  - Replace certain component of outlier detection by a DL model
  - Examples: OC-NN, DevNet





$$z = \frac{|x - \mu|}{\sigma}$$

Replace x by the output of a DL model

Summary

- Basic Concepts
  - Definition, types, challenges and applications
- Statistical Approaches
  - Parametric vs. non-parametric methods
- Proximity-Based Approaches
  - Distance based vs. density based
- Reconstruction-Based Approaches
  - Matrix factorization vs. pattern-based compression
- Outlier Detection in High-Dimensional Data
  - Subspace methods, ensemble, and deep learning

References (1)

- Charu C Aggarwal. Outlier analysis. In Data mining, pages 237– 263. Springer, 2015.
- Leman Akoglu, Hanghang Tong, and Danai Koutra. Graph based anomaly detection and description: a survey. Data mining and knowledge discovery, 29(3):626–688, 2015.
- Leman Akoglu, Hanghang Tong, Jilles Vreeken, and Christos Faloutsos. Fast and reliable anomaly detection in categorical data. In 21st ACM International Conference on Information and Knowl- edge Management, CIKM'12, Maui, HI, USA, October 29 - Novem- ber 02, 2012, pages 415–424, 2012.
- Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407, 2019.
- Raghavendra Chalapathy, Aditya Krishna Menon, and Sanjay Chawla. Anomaly detection using one-class neural networks. arXiv preprint arXiv:1802.06360, 2018.
- Manish Gupta, Jing Gao, Charu C Aggarwal, and Jiawei Han. Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and data Engineering, 26(9):2250–2267, 2013.
- Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. Deep learning for anomaly detection: A review. arXiv preprint arXiv:2007.02500, 2020.
- Hanghang Tong and Ching-Yung Lin. Non-negative residual matrix factorization with application to graph anomaly detection. In Pro- ceedings of the Eleventh SIAM International Conference on Data Mining, SDM 2011, April 28-30, 2011, Mesa, Arizona, USA, pages 143–153, 2011.

References (2)

- Hanghang Tong, Spiros Papadimitriou, Jimeng Sun, Philip S. Yu, and Christos Faloutsos. Colibri: fast mining of large static and dy- namic graphs. In Proceedings of the 14th ACM SIGKDD Interna- tional Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, pages 686–694, 2008.
- Jilles Vreeken, Matthijs van Leeuwen, and Arno Siebes. Krimp: mining itemsets that compress. Data Min. Knowl. Discov., 23(1):169–214, 2011.
- Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. IEEE transactions on information theory, 58(5):3047–3064, 2012.
- Chong Zhou and Randy C Paffenroth. Anomaly detection with ro- bust deep autoencoders. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Min- ing, pages 665–674, 2017.
- Jiong Zhang and Mohammad Zulkernine. Anomaly based network intrusion detection with unsupervised outlier detection. In 2006 IEEE International Conference on Communications, volume 5, pages 2388–2393. IEEE, 2006.
- Si Zhang, Dawei Zhou, Mehmet Yigit Yildirim, Scott Alcorn, Jin- grui He, Hasan Davulcu, and Hanghang Tong. Hidden: hierarchical dense subgraph detection with application to financial fraud detec- tion. SDM, 2017.