A Survey on Unsupervised Subspace Outlier Detection Methods for High-Dimensional Data

Jaehyeong Ahn

Department of Applied Statistics, Konkuk University

jayahn0104@gmail.com

June 3, 2021

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

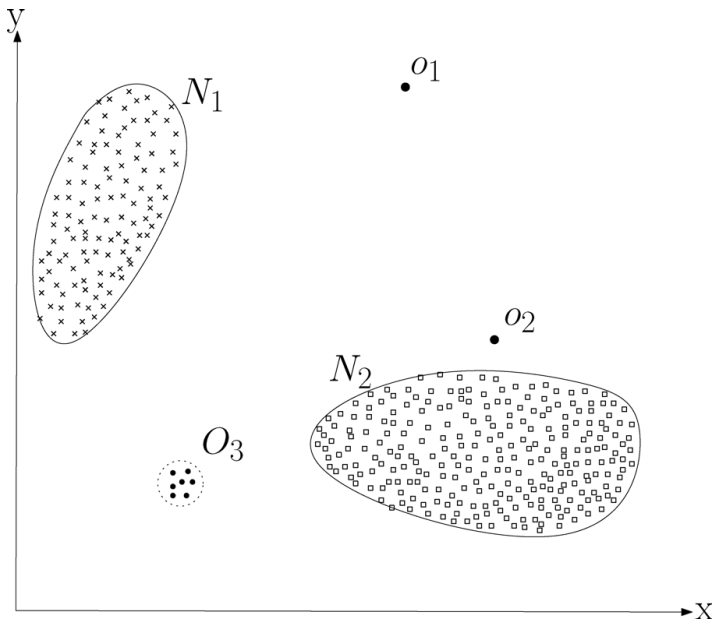
Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Rapid Introduction to Outlier Detection

What is an outlier?

- "Outlier is an observation that deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism" (Hawkins, 1980)



(Chandola, Banerjee, and Kumar, 2009)

Figure: Simple illustration of outliers in a two-dimensional space

Rapid Introduction to Outlier Detection

Why detecting outliers is important?

- In spite of the small numbers, outliers could cause siginificant and critical issues in the specific fields such as:
  - Credit card fraud detection (Fawcett and Provost, 1997)
  - Network intrusion detection (Eskin et al., 2002)
  - Medical diagnosis (Penny and Jolliffe, 2001)
  - Industrial damage detection (Basu and Meckesheimer, 2007)

Rapid Introduction to Outlier Detection

How to detect outliers?

- Outlier detection techniques can be catergorized based on the usage of the outlier label
  - Supervised Outlier Detection
    - Require full accurace class labels
    - Similar to solve an imbalanced classification problem
  - Semi-Supervised Outlier Detection
    - Require labels only for the normal instances
    - To build a model for the normal behavior, and use the model to identify outliers in the test data
  - Unsupervised Outlier Detection
    - Require no label information
    - Measure the degree of outlierness for each observation in the given data
    - Provide ranking with outlier score of each observation
- In this presentation, only the unsupervised outlier detection techniques will be considered

Rapid Introduction to Outlier Detection

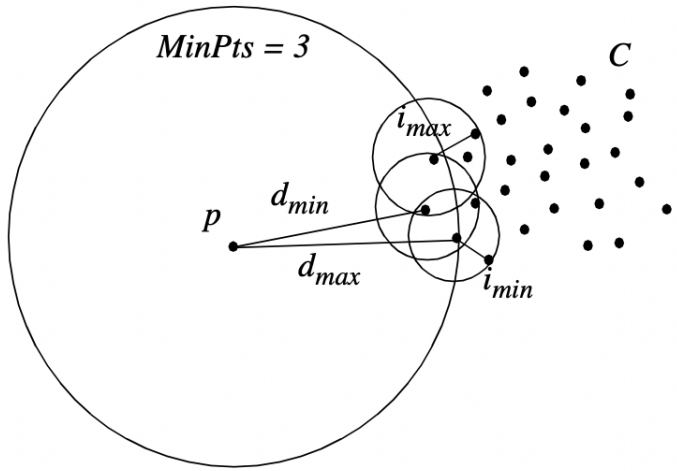
Representative Unsupervised Outlier Detection Techniques

- Probabilisitic Model-based Outlier Detection
  - Fit a specific probabilistic model to the given data
  - The degree of outlierness is defined as the inverse of the estimated probability density function
- Nearest Neighbor-based Outlier Detection
  - The degree of outlierness is calculated based on the specified distance function between observations in the feature space
- Clustering-based Outlier Detection
  - Fit a clustering algorithm to the given data and then compute the distance between the observations and their closest centroids

Rapid Introduction to Outlier Detection

Local Outlier Factor(LOF) (Breunig et al., 2000): The state of the art Nearest Neighbor-based unsupervised outlier detection technique

- LOF denotes the set of k nearest neighbors of an observation x_i as $N_k(x_i)$
- The local density of x_i is estimated by using the distance of the x_i to the k-th nearest neighbor of x_i
- The degree of outlierness is measured based on the ratio of the local density of x_i and the local densities of the observations in $N_k(x_i)$



(Breunig et al., 2000)

Figure: Simple illustration of LOF

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Concentration Effect

- Unsupervised Outlier Detection techniques generally measure the degree of outlierness by computing the distances between the observations in the feature space
- High-dimensional data usually causes a "concentration effect" which makes the distances between all observations become similar (Beyer et al., 1999)
- This phenomenon becomes the motivational problem of the outlier detection in high-dimensional data

Concentration Effect

Definition 1

- m: the number of dimensions
- F_{data_m}: an infinite sequence of data distributions, $m = 1, 2, \cdots$
  - $X^{(m)} \sim F_{data_m}$: an arbitrary random vector distributed as F_{data_m}
  - $x_1^{(m)}, \cdots, x_n^{(m)} \sim F_{data_m}$: n independent data points per m
- F_{query_m}: an infinite sequence of query distributions, $m = 1, 2, \cdots$
  - $Q^{(m)} \sim F_{query_m}$: a query point chosen independently from $x_i^{(m)}, \forall_i$
- $d_{m,p}(X^{(m)}, Q^{(m)})$: the function gives the L_p distance between $X^{(m)}$ and $Q^{(m)}$, $\forall p > 0$
- $\mathrm{DMAX}^{(m)} = \max\{d_{m,p}(x_i^{(m)}, Q^{(m)}) | 1 \leq i \leq n\}$
- $\mathrm{DMIN}^{(m)} = \min\{d_{m,p}(x_i^{(m)}, Q^{(m)}) | 1 \leq i \leq n\}$

Concentration Effect

- Below theorem states that assuming the distance distribution behaves a certain way as m increases, *the difference in distance between the query point and all data points becomes "negligible"*

Theorem 1 (*ConcentrationEffect*)

Under the certain conditions in Definition 1,

If $\lim_{m\to\infty} \mathrm{Var}\left(\frac{\mathrm{d_{m,p}(X^{(m)},Q^{(m)})}}{\mathrm{E[d_{m,p}(X^{(m)},Q^{(m)})]}}\right) = 0$, Then $\frac{\mathrm{DMAX^{(m)}}}{\mathrm{DMIN^{(m)}}} \to_p 1$

  - Where the operators $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ refer to the theoretical expectation and variance of the distribution of the random variable $d_{m,p}(X^{(m)}, Q^{(m)})$
  - It is assumed that $\mathrm{E[d_{m,p}(X^{(m)}, Q^{(m)})]}$ and $\mathrm{Var(d_{m,p}(X^{(m)}, Q^{(m)}))}$ are finite and $\mathrm{E[d_{m,p}(X^{(m)}, Q^{(m)})]} \neq 0$

Concentration Effect

Fundamental Questions

1. When does the condition of Theorem 1 hold?

   (i.e., $\lim_{m\to\infty} \mathrm{Var}\left(\frac{\mathrm{d_{m,p}(X^{(m)},Q^{(m)})}}{\mathrm{E[d_{m,p}(X^{(m)},Q^{(m)})]}}\right) = \lim_{\mathrm{m}\to\infty} \frac{\mathrm{Var(d_{m,p}(X^{(m)},Q^{(m)}))}}{(\mathrm{E[d_{m,p}(X^{(m)},Q^{(m)})]})^2} = 0$)

2. For situations in which the condition is satisfied, "at what rate" do distances between observations become indistinct as dimensionality increases?

- To answer these questions, a set of simulations will be provided

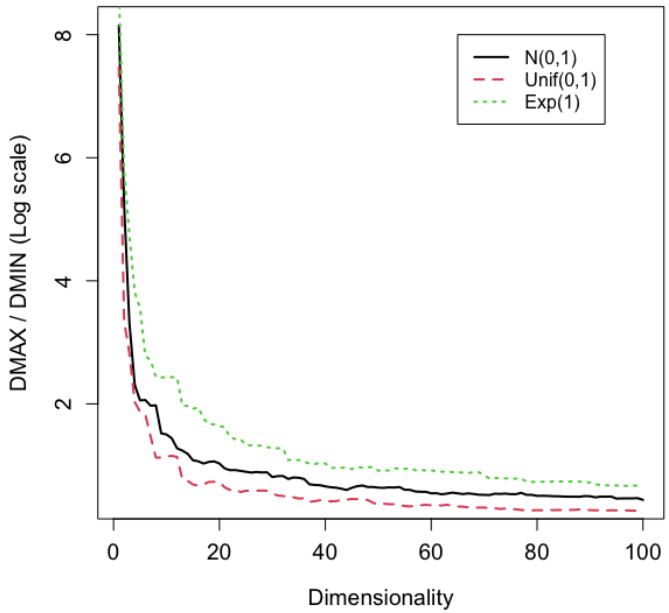
Concentration Effect

Empirical experiments

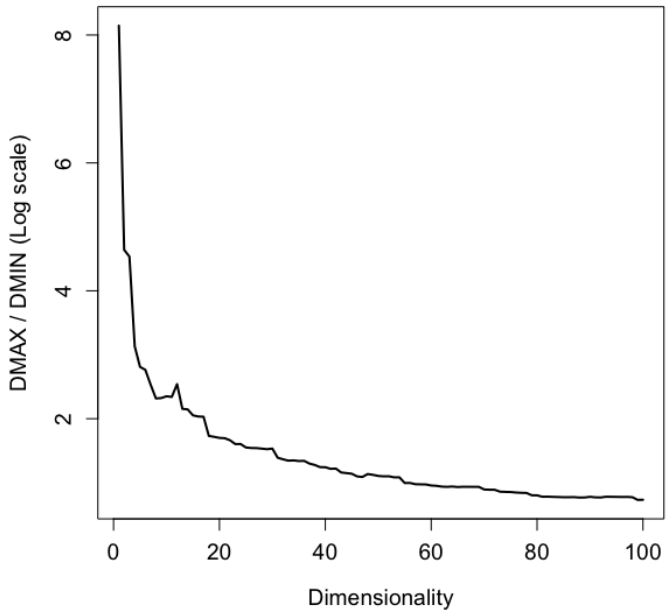
- Case 1) Independent and Identically Distributed (IID) Dimensions
  - Each marginal distribution is independent and identically distributed(i.i.d.) in all dimensions
  - For example, $X_j \sim N(0,1),\ \forall j$
- Case 2) Marginal Data Distributions change with Dimensionality
  - The marginal distributions of data change with dimensionality
  - For instance, $X_1 \sim Unif(0,1), X_2 \sim N(0,1), X_3 \sim Exp(1)$

Empirical experiments

- Below figures show the relationship between dimension size and DMAX/DMIN with various distributions



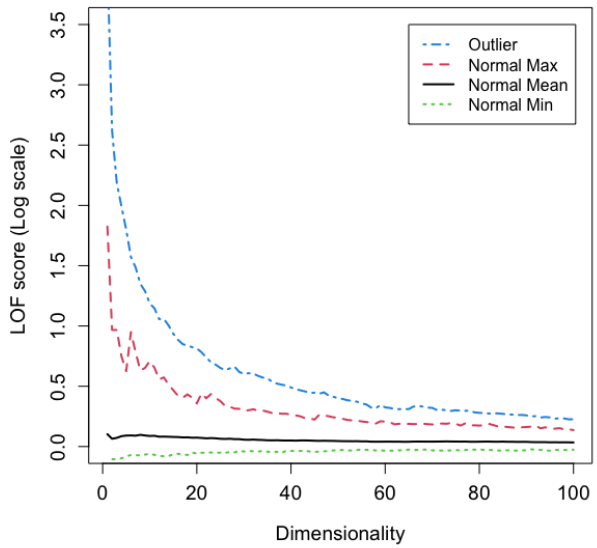
(a) Case 1) IID dimensions (N(0,1), Unif(0,1), Exp(1)) - 1K samples



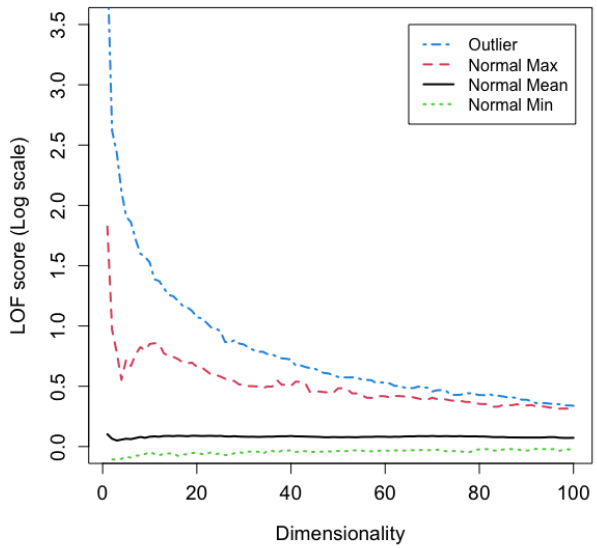
(b) Case 2) - Randomly chosen marginal distributions (N(0,1), Unif(0,1), Exp(1)) - 1K samples

Empricial experiments: With Outlier

- The data generated earlier were inherited, but the first observatoin x_1 becomes an outlier by setting $x_{11} = 10$
- The degree of outlierness measured by applying LOF (wtih k =10) according to the dimension size is provided in the below figures



(a) Case 1) IID dimensions following N(0,1) with outlier x_1- 1K samples

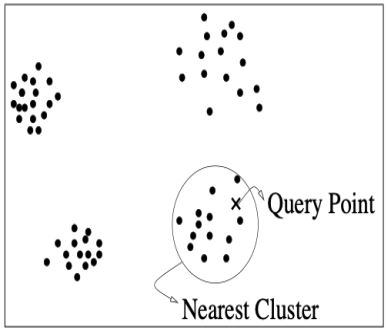


(b) Case 2) Randomly chosen marginal distributions (N(0,1), Unif(0,1), Exp(1)) with outlier x_1- 1K samples

Concentration Effect

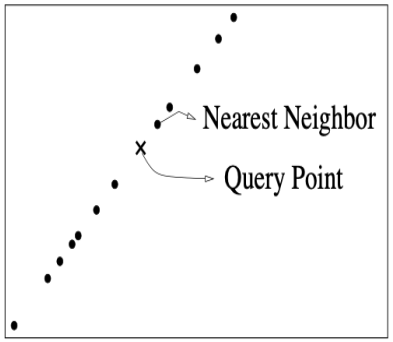
Generalized Cases that can Prevent Concentration Effect

1. Separable Cluster Structure
   - Bennett et. al. (1999) proved that when there exist separable cluster structure, the distance between two data points in different clusters (between cluster distance) dominates the distance between two points in the same cluster (within cluster distance).
2. Low Intrinsic Structure
   - Durrant and Kaban (2009) showed that when there exist richness of correlations between the variables, the concentration phenomenon would not appear



(Beyer et al., 1999)

(a) Separable Cluster Structure

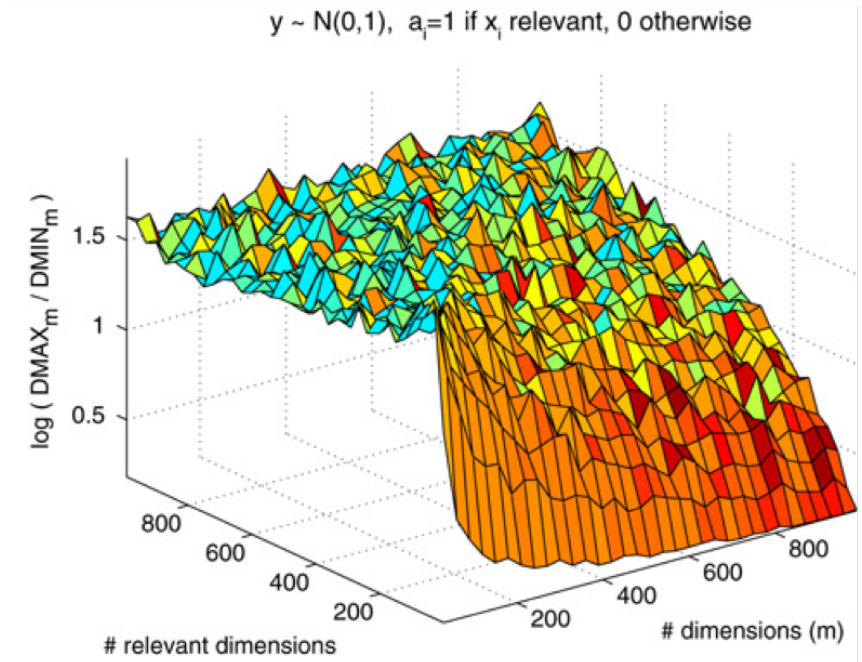


(Beyer et al., 1999)

(b) Low Intrinsic Structure

Concentration Effect

- Durrant and Kaban (2004) defines the "relevant dimension" as the variable which has correlation with other variables in the perspective of the latent variable model
- Below Results show that the proportion of relevant dimensions among all dimensions plays a key role for preventing the concentration effect



(Durrant and Kabán, 2009)

Figure: Relationship between the portion of relevant dimensions and DMAX/DMIN

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Subspace Outlier Detection

- Subspace outlier detection overcomes the concentration effect in high-dimensional data by:
  - Selecting subspaces (i.e., sets of variables) that satisfy a specific criterion, such as having a cluster structure or having a strong correlation
  - Measuring the degree of outlierness for each observation in that subspaces
- In this presentation, seven representative methodologies will be briefly introduced according to the specific criteria (detailed explanations are provided in the thesis paper)

Subspace Outlier Detection

- All methodologies are categorized into three types: Random subspace search, Local subspace search, Global subspace search according to the subspace selection method

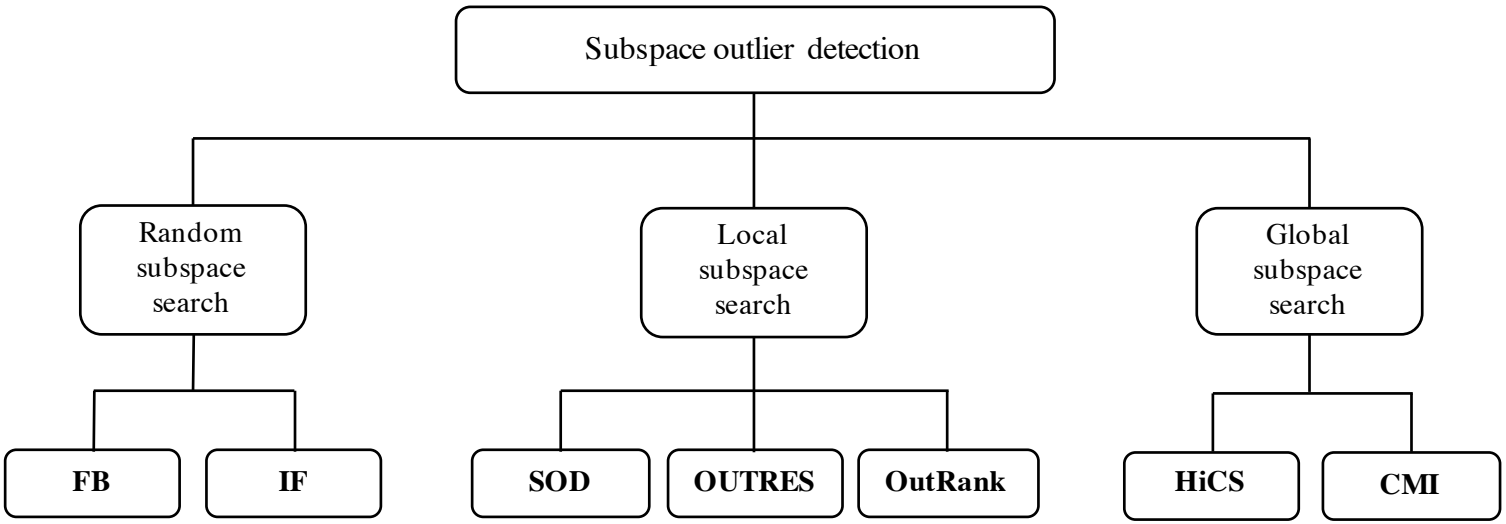


Figure: Taxonomy of subspace outlier detection techniques

Subspace Outlier Detection

- Each methodology will be investigated according to the subspace search, the subspace result, and the outlierness measurement

Method	Subspace search		Subspace result		Outlierness function		Implementation tool
	Criteria	Algorithm	Local vs Global	Single vs Multiple	Subspace outlierness	Combination	
FB	Random	None	Global	Multiple	$\mathrm{COD}_{S^k}(x_i)$	Sum	ELKI
IF	Random	None	Global	Multiple	$\mathrm{PathLength}_{T^k}(x_i)$	Average	R
SOD	Distance from mean on one diemensional space	Brute force	Local	Single	$\mathrm{SOD}(x_i)$	None	ELKI
OUTRES	Goodness of fit test for uniform distribution	Apriori	Local	Multiple	$\mathrm{Score}_S(x_i)$	1 - Product	ELKI
OutRank	Depends on the subspace clustering	Depends on the subspace clustering	Local	Multiple	$\mathrm{Evid}_k(x_i)$	1-Average	ELKI
HiCS	Contrast	Beam search	Global	Multiple	$\mathrm{COD}_S(x_i)$	Average	ELKI
CMI	CMI	Beam search	Global	Multiple	$\mathrm{COD}_S(x_i)$	Average	Java

Figure: Summary table of subspace outlier detection methods

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Random Subspace Search

- Random subspace search method measures the degree of outlierness of the observations in randomly selected subspaces
- In order to compensate for the randomness of subspace selection, this technique summarizes the outliernesses calculated in several subspaces with a combination function such as an average

Random Subspace Search: FB

Feature Bagging (FB) (Lazarevic and Kumar, 2005)

- **FB** ramdomly selects m subspaces and then applying the existing classical outlier detection technique (e.g., LOF) to measure the degree of outlierness and then summarize the outliernesses in all subspaces

Random Subspace Search: FB

(A) Determine user-defined elements

- COD: Classical outlier detection technique to be used in subspace
- m: The number of subspaces to be selected

(B) Select a set of subspaces

$$\text{SS} = \{\mathcal{S}^k : k \leq m\}$$

- $\mathcal{S}^k$: Any subspace that satisfies $d/2 \leq |\mathcal{S}^k| \leq d-1$

(C) Measure the degree of outlierness of the observation

$$\text{FB}(x_i) = \text{Sum}(\{\text{COD}_{\mathcal{S}^k}(x_i) : \mathcal{S}^k \in \text{SS}\})$$

- $\text{COD}_{\mathcal{S}^k}(x_i)$: The subspace outlierness measured using COD in subspace $\mathcal{S}^k$

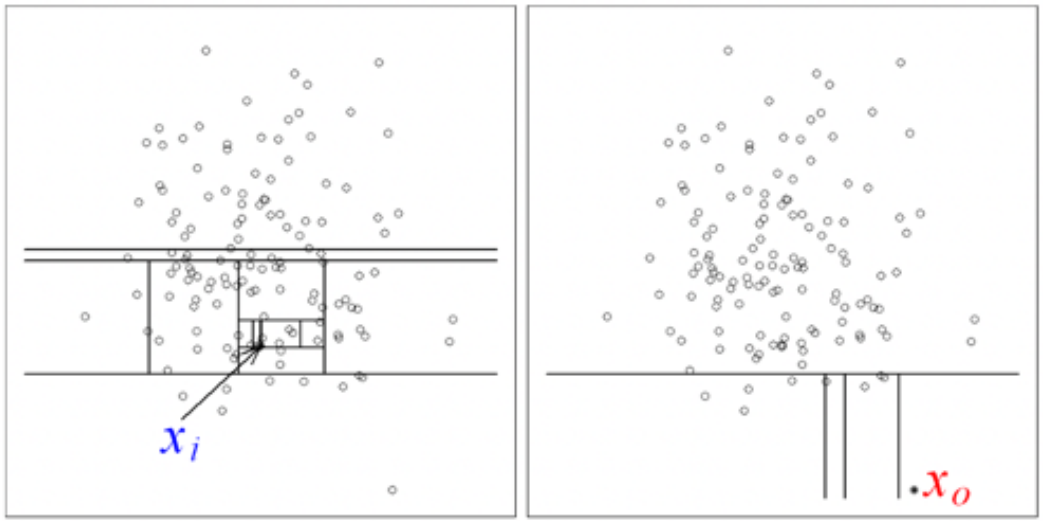
Random Subspace Search: IF

Isolation Forest (IF) (Liu, Ting, and Zhou, 2008)

- **IF** creates m unsupervised decision trees which randomly partition the feature space with binary random split until every single observation is isolated (i.e., every observation is in the different terminal nodes)
- The degree of outlierness in each decision tree is calculated based on the number of splits to be isolated
- The final degree of outlierness is summarized by averaging the outlierness in all decision trees

Random Subspace Search: IF

- It is based on the assumption that normal observations are in a dense area, so that the number of splits required for isolation is large, while outliers are separted from most observations so that the number of splits required for isolation is relatively small



(Chandola, Banerjee, and Kumar, 2009)

Figure: Simple illustration of IF

Random Subspace Search: IF

(A) Determine user-defined elements

- m: The number of decision trees

(B) Create a decision tree that isolates all observations

(a) Create a bootstrapped sample without replace ment X^k

(b) Construct a decision tree T^k from X^k by applying the following rules

* Splitting rule: Use an arbitrary variable X_j and an arbitrary threshold $c \in [\min(X_j), \max(X_j)]$
* Stopping rule: Stop if the number of observations in all terminal nodes is 1

(C) Calculate the degree of outlierness of the observation

$$\text{IF}(x_i) = \text{Average}(\{\text{PathLength}_{T^k}(x_i) : k \leq m\})$$

- $\text{PathLength}_{T^k}(x_i)$: The depth of terminal node in which x_i belongs to in T^k

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Local Subspace Search

- Local subspace search method measures the degree of outlierness by selecting relevant subspaces for each observation
- If there are multiple selected subspaces, the final outlierness is defined using an appropriate combination function

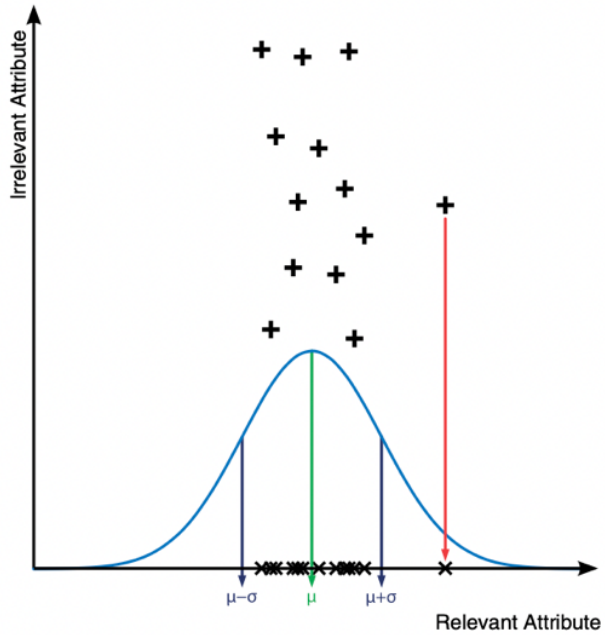
Local Subspace Search: SOD

Subspace Outlier Degree (SOD) (Kriegel et al., 2009)

- **SOD** selects one relevant subspace for each observation to measure the degree of outlierness
- It tries to find variables in which the nearest neighbors of an observation are clustered in its one-dimensional feature space

Local Subspace Search: SOD

- Subspace selection
  - It determines the set of shared nearest neighbors of x_i as SN(x_i)
  - By projecting SN(x_i) onto each variable, the clustered variables are adopted as relevant variables
  - The set of relevant variables is obtained as relevant subspace of x_i
- Outlierness measurement
  - The degree of outlierness of x_i is measured based on the distance from SN(x_i) on the relevant subspace of x_i



(Zimek, Schubert, and Kriegel, 2012)

Figure: Simple illustration of SOD

Local Subspace Search: SOD

(A) Determine user-defined elements

- k: The number of nearest neighbors
- l: The number of shared nearest neighbors
- α: The ratio for squared distance

(B) Determine the set of shared nearest neighbors.

$$\mathrm{SN}(x_i) = \left\{x_t : S_k(x_i, x_t) \geq \mathrm{Upper}_\ell(\{S_k(x_i, x_s) : s \neq i\})\right\},$$

- $S_k(x_i, x_t) = \left|N_k(x_i) \cap N_k(x_t)\right|$
- $N_k(x_i) = \left\{x_t : \mathrm{Dist}_{\mathcal{D}}(x_i, x_t) \leq \mathrm{Lower}_k(\{\mathrm{Dist}_{\mathcal{D}}(x_i, x_s) : s \neq i\})\right\}$

(C) Select the relevant subspace of x_i

$$\mathcal{S}_\alpha(x_i) = \left\{j : \sigma^2_{\{j\}}(x_i) < \alpha\sigma^2_{\mathcal{D}}(x_i)/d, j \in \mathcal{D}\right\},$$

- $\sigma^2_{\{j\}}(x_i) = \mathrm{Average}\big(\{\mathrm{Dist}_{\{j\}}(x_t, \mu(x_i))^2 : x_t \in \mathrm{SN}(x_i)\}\big)$
- $\sigma^2_{\mathcal{D}}(x_i) = \mathrm{Average}\big(\{\mathrm{Dist}_{\mathcal{D}}(x_t, \mu(x_i))^2 : x_t \in \mathrm{SN}(x_i)\}\big)$
- $\mu(x_i) = \mathrm{Average}\big(\{x_t : x_t \in \mathrm{SN}(x_i)\}\big)$

(D) Calculate the degree of outlierness of the observation

$$\mathrm{SOD}(x_i) = |\mathcal{S}_\alpha(x_i)|^{-1}\mathrm{Dist}_{\mathcal{S}_\alpha(x_i)}\big(x_i, \mu(x_i)\big)$$

Local Subspace Search: OUTRES

OUTlier ranking in RElevant Subspaces (OUTRES) (Müller, Schiffer, and Seidl, 2011)

- **OUTRES** selects multiple relevant subspaces for each observation
- It tries find the subspaces that have some cluster structures for the set of nearest neighbors of an observation in a specific subspace

Local Subspace Search: OUTRES

- Subspace selection
  - Given a subspace $\mathcal{S} \subset \mathcal{D}$, a nearest neighbor set of x_i, $N_{\mathcal{S},\epsilon}(x_i)$ is determined
  - On $N_{\mathcal{S},\epsilon}(x_i)$ in $\mathcal{S}$, the Kolmogorove-Smirnov goodness of fit test for Uniform distribution is conducted. If the test rejects the null-hypothesis, $\mathcal{S}$ is obtained as a relevant subspace
  - However, it requires $2^d - 1$ computational cost to apply the above test to all possible combinations of subspaces. Thus an apriori-like algorithm is proposed
- Outlierness measurement
  - Given a relevant subspace $\mathcal{S}$, the degree of normality of x_i is measured based on the local density and the local deviation of x_i
  - The final degree of outlierness of x_i is calculated combining the scores in all relevant subspaces of x_i

Local Subspace Search: OUTRES

(A) Determine the user-defined elements

- ε: Neighborhood range setting constant
- $\alpha \in (0, 1)$: Significance level for the test

(B) Search the set of relevant subspaces of x_i

$$\mathrm{SS}(x_i) = \{\mathrm{SS}^k(x_i) : k \leq d\}$$

- $\mathrm{SS}^1(x_i) = \left\{\mathcal{S} : \mathrm{KS}_{\mathcal{S},\alpha}(x_i) = 1, |\mathcal{S}| = 1\right\}$
- $\mathrm{SS}^k(x_i) = \left\{\mathcal{S} \cup \mathcal{S}' : \mathrm{KS}_{\mathcal{S}\cup\mathcal{S}',\alpha}(x_i) = 1, |\mathcal{S} \cup \mathcal{S}'| = k,\ \mathcal{S}, \mathcal{S}' \in \mathrm{SS}^{k-1}(x_i)\right\}, 2 \leq k \leq d$

(C) Measure the degree of normality of x_i in $\mathcal{S} \in \mathrm{SS}(x_i)$

$$\mathrm{Score}_{\mathcal{S}}(x_i) = \begin{cases} \mathrm{Den}_{\mathcal{S}}(x_i)/\mathrm{Dev}_{\mathcal{S}}(x_i), & \text{if } \mathrm{Dev}_{\mathcal{S}}(x_i) \geq 1 \\ 1, & \text{else} \end{cases}$$

- $\mathrm{Den}_{\mathcal{S}}(x_i) = \mathrm{Average}\left(\left\{K\left(\mathrm{Dist}_{\mathcal{S}}(x_i, x_t)\right)/v_{\varepsilon}(|\mathcal{S}|) : x_t \in N_{\mathcal{S},\varepsilon}(x_i)\right\}\right)$
- $\mathrm{Dev}_{\mathcal{S}}(x_i) = \left(\mu_{\mathcal{S}}(x_i) - \mathrm{Den}_{\mathcal{S}}(x_i)\right)/2\sigma_{\mathcal{S}}(x_i)$
  - $\sigma_{\mathcal{S}}(x_i) = \mathrm{Average}\left(\left\{\left(\mathrm{Den}_{\mathcal{S}}(x_t) - \mu_{\mathcal{S}}(x_i)\right)^2 : x_t \in N_{\mathcal{S},\varepsilon}(x_i)\right\}\right)^{1/2}$
  - $\mu_{\mathcal{S}}(x_i) = \mathrm{Average}\left(\left\{\mathrm{Den}_{\mathcal{S}}(x_t) : x_t \in N_{\mathcal{S},\varepsilon}(x_i)\right\}\right)$

(D) Calculate the degree of outlierness of observation

$$\mathrm{OUTRES}(x_i) = 1 - \mathrm{Product}(\{\mathrm{Scroe}_{\mathcal{S}}(x_i) : \mathcal{S} \in \mathrm{SS}(x_i)\})$$

Local Subspace Search: OutRank

Outlier Ranking via subspace analysis (OutRank) (Müller et al., 2012)

- **OutRank** applies the subspace clustering method directly for subspace outlier detection by obtaining relevant subspaces for each observation based on the results of a subspace clustering algorithm
- Subspace clustering explores clusters embedded in subspaces of the given data and then outputs the set of subspace clusters and according subspaces

Local Subspace Search: OutRank

- Subspace selection
  - Get a set of subspace clustering results by fitting a specific subspace clustering algorithm to the given data
    - The clustering algorithm does not assign cluster membership to some observations
  - Only the subspaces in which x_i belongs to a subspace cluster are obtained as relevant subspaces
- Outlierness measurement
  - The degree of normality of x_i is measured based on the size of the subspace cluster which x_i belongs to, and the size of subspace dimension
  - The final degree of outlierness is calculated combining the scores in all relevant subspaces of x_i

Local Subspace Search: OutRank

(A) Determine the user-defined elements

- SC: A subspace clustering technique

(B) Generate a subspace cluster set of data by using SC

$$\mathrm{SCR} = \left\{(C_k, \mathcal{S}_k) : k \leq m\right\}$$

- m: The number of subspaces
- C_k: A cluster set configured in the subspace $\mathcal{S}_k$

(C) Measure the degree of normality of x_i

$$\mathrm{Regular}(x_i) = \mathrm{Average}\big(\{\mathrm{Evid}_k(x_i) : k \in \mathrm{SCR}(x_i)\}\big)$$

- $\mathrm{SCR}(x_i) = \{k : |C_k(x_i)| \neq 0, k \leq m\}$
- $C_k(x_i)$: A cluster included in C_k to which x_i belongs
- $\mathrm{Evid}_k(x_i) = |C_k(x_i)|/\max_{j\leq m}|C_j| + |\mathcal{S}_k|/\max_{j\leq m}|\mathcal{S}_j|$

(D) Calculate the degree of outlierness

$$\mathrm{OutRank}(x_i) = 1 - \mathrm{Regular}(x_i)$$

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Global Subspace Search

- Global subspace search method explores for relevant subspaces that are suitable for all observations in the given data
- The final degree of outlierness is calucated by combining the outlierness in all the relevant subspaces
- This technique can be considered as a preprocessing step for outlier detection cause the subspace selection stage and the outlierness measurement stage are separated so that the existing outlier detection method can be borrowed for measuring the degree of outlierness

Global Subspace Search: HiCS

High Contrast Subspaces (HiCS) (Keller, Muller, and Bohm, 2012)

- **HiCS** tries to select subspaces with strong mutual correlations among variables as relevant subspaces for the given data
- For this purpose, **HiCS** proposes a function "Contrast" that can measure the degree of mutual correlation between variables in over two-dimensional space

Global Subspace Search: HiCS

Definition (Contrast)

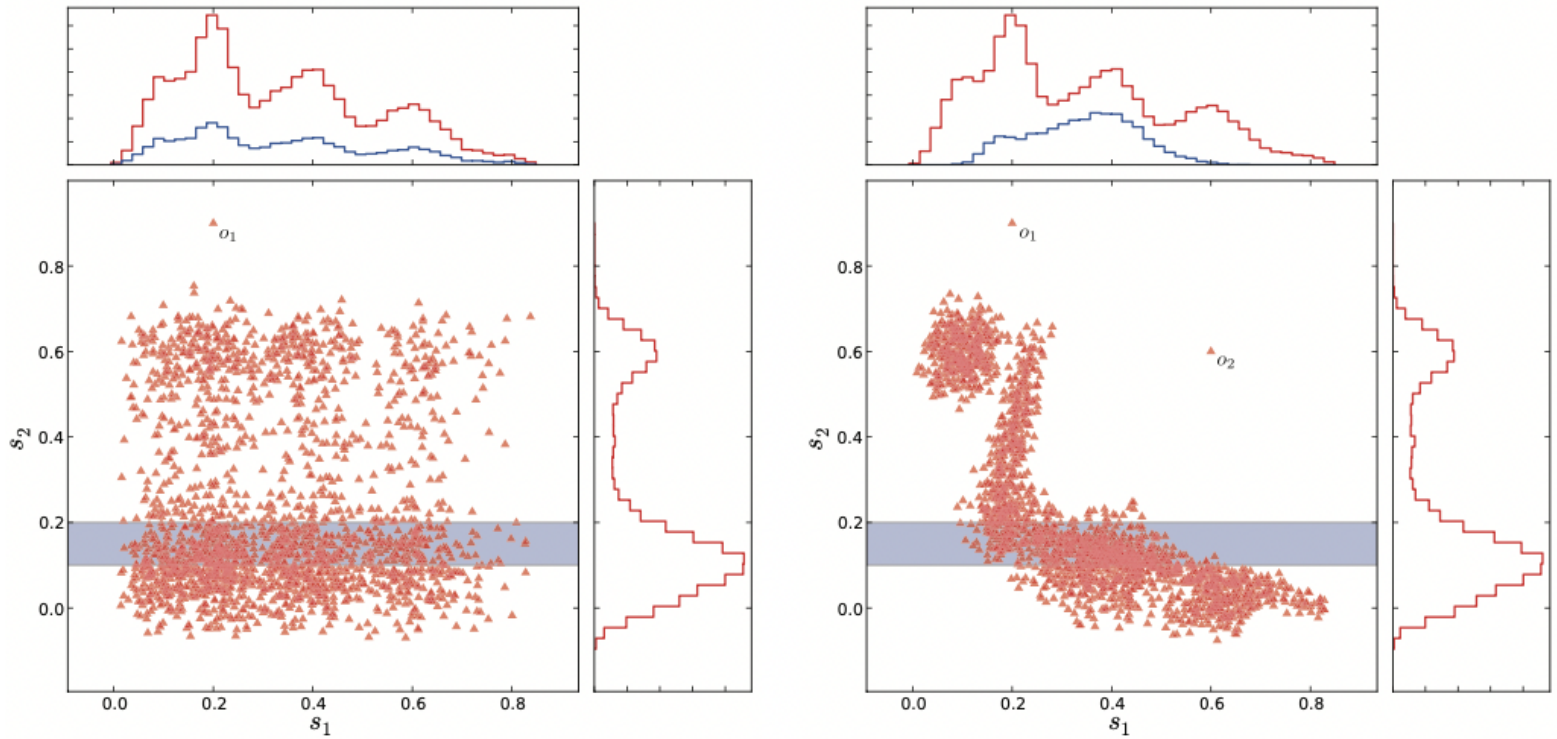
$$\text{Contrast}(\mathcal{S}) = \text{Average}\big(\{\text{Dev}(X_j, X_j|R_j) : j \in \mathcal{S}'\}\big), \mathcal{S} \subset \mathcal{D}, |\mathcal{S}| \geq 2$$

- $\mathcal{S}' = \{j_1, \ldots, j_m\}$: A set of m variable indexes sampled with replacement from $\mathcal{S}$
- $R_j = \prod_{l \in \mathcal{S}, l \neq j} [a_l, b_l]$: A rectangular subset consisting of any section $[a_l, b_l]$ which satisfies the following condition

$$\text{Average}\big(\{I(x_{il} \in [a_l, b_l]) : i \leq n\}\big) \geq \alpha^{-1/|\mathcal{S}|}, \ \alpha \in (0, 1)$$

- $\text{Dev}(X_j, X_j|R_j) = \sup_{x_{ij} \in X_j|R_j} \left|F_{X_j}(x_{ij}) - F_{X_j|R_j}(x_{ij})\right|$
  - $X_j|R_j = \{x_{ij} : x_{il} \in [a_l, b_l], l \in \mathcal{S}, l \neq j\}$
  - $F_{X_j}(x_{ij}) = \text{Average}\big(\{I(x_{sj} < x_{ij}) : x_{sj} \in X_j\}\big)$
  - $F_{X_j|R_j}(x_{ij}) = \text{Average}\big(\{I(x_{sj} < x_{ij}) : x_{sj} \in X_j|R_j\}\big)$

Global Subspace Search: HiCS



(Keller, Muller, and Bohm, 2012)

Figure: Graphical illustration Contrast

Global Subspace Search: HiCS

- Subspace selection
  - For a given subspace $\mathcal{S} \subset \mathcal{D}$, Compute the Contrast($\mathcal{S}$)
  - But, it requires expensive computational cost to apply the Contrast function to all possible combination of subspaces. Thus the beam-search algorithm is applied
  - Only the subspaces with high Contrast values are obtained as relevant subspaces of the given data
- Outlierness measurement
  - Given a relevant subspace $\mathcal{S}$, the degree of outlierness of all observation is measured by applying classical outlier detection technique (e.g., LOF)
  - The final degree of outlierness of all observation is computed by averaging the scores in all relevant subspaces

Global Subspace Search: HiCS

(A) Determine the user-defined elements

- COD: Classical outlier detection technique to be used in subspace
- $\alpha \in (0, 1)$: The ratio of data in the interval $[a_l, b_l]$ in the Contrast function
- m: The number of Dev functions used for Contrast measurement
- t: The number of subspaces selected for each subspace search step

(B) Explore a set of relevant subspaces of data

$$\mathrm{SS} = \{\mathrm{SS}^k : 2 \leq k \leq d\}$$

- $\mathrm{SS}^2 = \{\mathcal{S} : \mathrm{Contrast}(\mathcal{S}) \geq \mathrm{Upper}_t(\{\mathrm{Contrast}(\mathcal{S}) : |\mathcal{S}| = 2\}), |\mathcal{S}| = 2\}$
- $\mathrm{SS}^k = \{\mathcal{S} \cup \mathcal{S}' : \mathrm{Contrast}(\mathcal{S} \cup \mathcal{S}') \geq \mathrm{Upper}_t(\{\mathrm{Contrast}(\mathcal{S} \cup \mathcal{S}') : |\mathcal{S} \cup \mathcal{S}'| = k, \mathcal{S}, \mathcal{S}' \in \mathrm{SS}^{k-1}\}), |\mathcal{S} \cup \mathcal{S}'| = k\}, 3 \leq k \leq d$
- For $\mathcal{S}, \mathcal{S}' \in \mathrm{SS}$, remove $\mathcal{S}'$ which $\mathcal{S}' \subset \mathcal{S}$ and $\mathrm{Contrast}(\mathcal{S}') < \mathrm{Contrast}(\mathcal{S})$

(C) Calculate the degree of outlireness of the observation

$$\mathrm{HiCS}(x_i) = \mathrm{Average}\big(\{\mathrm{COD}_{\mathcal{S}}(x_i) : \mathcal{S} \in \mathrm{SS}\}\big)$$

- $\mathrm{COD}_{\mathcal{S}}(x_i)$: Subspace outlierness calculated by COD in the subspace $\mathcal{S}$

Global Subspace Search: CMI

Cumulative Mutual Information (CMI) (Nguyen et al., 2013)

- **CMI** is an advanced technique of HiCS, proposing a novel measure "CMI" to measure the correlation among variables in a subspace
- "CMI", unlike "Contrast", excludes the randomness of computation and uses cumulative entropy for calculation instead of cumulative distribution function
- Other than that, the procedures are identical to HiCS

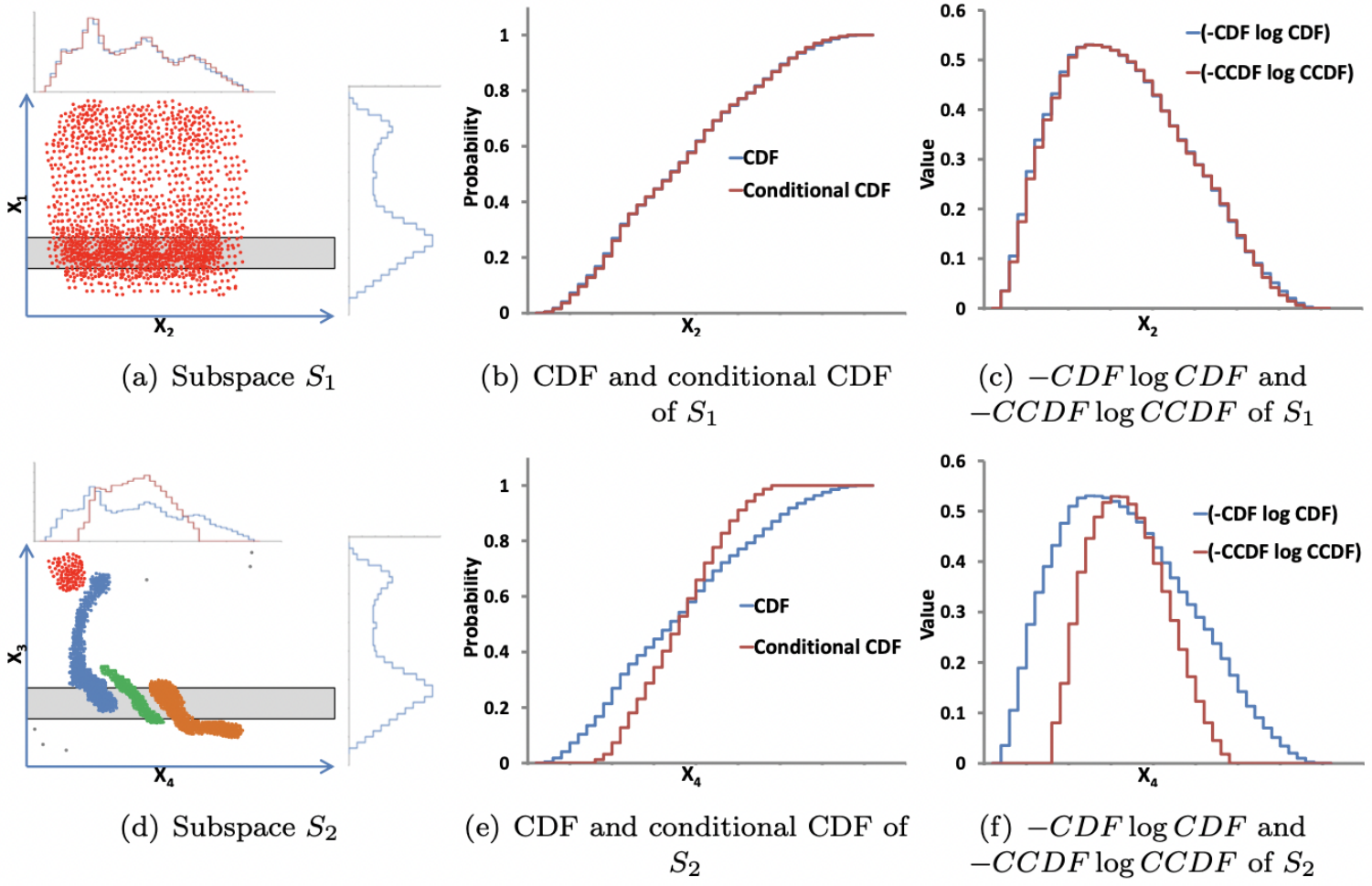
Global Subspace Search: CMI

Definition (CMI)

$$\mathsf{CMI}(\mathcal{S}) = \mathsf{Average}\big(\{\mathsf{Diff}(X_{j_k}, X_{j_k}|\mathsf{Clust}(X_{\mathcal{S}_{k-1}})) : 2 \leq k \leq |\mathcal{S}|\}\big), \mathcal{S} \subset \mathcal{D}$$

- $\mathsf{Clust}(X_t) = \{C_{t1}, \ldots, C_{tm_t}\}$: A set of m_t clusters configured on X_t
- $\mathsf{Diff}(X_s, X_s|\mathsf{Clust}(X_t)) = \mathsf{CumEnt}(X_s) - \sum_{l=1}^{m_t}(|C_{tl}|/n)\mathsf{CumEnt}(X_s|C_{tl})$
  - $\mathsf{CumEnt}(X_s) = -\sum_{i \leq |X_s|-1}(x_{(i+1)s} - x_{(i)s})(i/|X_s|)\log(i/|X_s|)$
    - $x_{(i)s} = \mathsf{Upper}_i(X_s)$
  - $\mathsf{CumEnt}(X_s|C_{kl}) = -\sum_{i \leq |X_s|C_{kl}|-1}(x_{(i+1)s} - x_{(i)s})(i/|X_s|C_{kl}|)$ $\log(i/|X_s|C_{kl}|)$
    - $X_s|C_{tl} = \{x_{is} : i \in C_{tl}\}$

Global Subspace Search: CMI



(a) Subspace S_1

(b) CDF and conditional CDF of S_1

(c) $-CDF \log CDF$ and $-CCDF \log CCDF$ of S_1

(d) Subspace S_2

(e) CDF and conditional CDF of S_2

(f) $-CDF \log CDF$ and $-CCDF \log CCDF$ of S_2

(Keller, Muller, and Bohm, 2012)

Figure: Graphical illustration CMI

Global Subspace Search: CMI

- Subspace selection
  - For a given subspace $\mathcal{S} \subset \mathcal{D}$, Compute the CMI($\mathcal{S}$)
  - But, it requires expensive computational cost to apply the CMI function to all possible combination of subspaces. Thus the beam-search algorithm is applied
  - Only the subspaces with high CMI values are obtained as relevant subspaces of the given data
- Outlierness measurement
  - Given a relevant subspace $\mathcal{S}$, the degree of outlierness of all observation is measured by applying classical outlier detection technique (e.g., LOF)
  - The final degree of outlierness of all observation is computed by averaging the scores in all relevant subspaces

Global Subspace Search: CMI

(A) Determine user-defined elements

- COD: Classical outlier detection technique to be used in subspace
- Clust: Clustering technique to be used in subspace
- t: The number of subspaces selected for each subspace search step

(B) Explore a set of relevant subspaces of data

$$\mathrm{SS} = \{\mathrm{SS}^k : 2 \leq k \leq d\}$$

- $\mathrm{SS}^2 = \{\mathcal{S} : \mathrm{CMI}(\mathcal{S}) \geq \mathrm{Upper}_t(\{\mathrm{CMI}(\mathcal{S}) : |\mathcal{S}| = 2\}), |\mathcal{S}| = 2\}$
- $\mathrm{SS}^k = \{\mathcal{S} \cup \mathcal{S}' : \mathrm{CMI}(\mathcal{S} \cup \mathcal{S}') \geq \mathrm{Upper}_t(\{\mathrm{CMI}(\mathcal{S} \cup \mathcal{S}') : |\mathcal{S} \cup \mathcal{S}'| = k,\ \mathcal{S}, \mathcal{S}' \in \mathrm{SS}^{k-1}\}), |\mathcal{S} \cup \mathcal{S}'| = k\}, 3 \leq k \leq d$
- For $\mathcal{S}, \mathcal{S}' \in \mathrm{SS}$, remove $\mathcal{S}'$ which $\mathcal{S}' \subseteq \mathcal{S}$ and $\mathrm{CMI}(\mathcal{S}') < \mathrm{CMI}(\mathcal{S})$

(C) Calculate the degree of outlireness of the observation

$$\mathrm{CMI}(x_i) = \mathrm{Average}\big(\{\mathrm{COD}_{\mathcal{S}}(x_i) : \mathcal{S} \in \mathrm{SS}\}\big)$$

- $\mathrm{COD}_{\mathcal{S}}(x_i)$: Subspace outlierness calculated by COD on subspace $\mathcal{S}$

Subspace Outlier Detection

Method	Subspace search		Subspace result		Outlierness function		Implementation tool
	Criteria	Algorithm	Local vs Global	Single vs Multiple	Subspace outlierness	Combination	
FB	Random	None	Global	Multiple	$\text{COD}_{S^k}(x_i)$	Sum	ELKI
IF	Random	None	Global	Multiple	$\text{PathLength}_{T^k}(x_i)$	Average	R
SOD	Distance from mean on one diemensional space	Brute force	Local	Single	$\text{SOD}(x_i)$	None	ELKI
OUTRES	Goodness of fit test for uniform distribution	Apriori	Local	Multiple	$\text{Score}_S(x_i)$	1 - Product	ELKI
OutRank	Depends on the subspace clustering	Depends on the subspace clustering	Local	Multiple	$\text{Evid}_k(x_i)$	1-Average	ELKI
HiCS	Contrast	Beam search	Global	Multiple	$\text{COD}_S(x_i)$	Average	ELKI
CMI	CMI	Beam search	Global	Multiple	$\text{COD}_S(x_i)$	Average	Java

Figure: Summary table of subspace outlier detection methods

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Evaluation of Unsupervised Outlier Detection

- The evaluation of outlier detection generally follows the evaluation method of binary classification
- However, in the case of unsupervised outlier detection, it is difficult to directly use binary classification evaluation measures since it is hard to set a clear threshold to determine wheter or not an observation is an outlier

Evaluatoin of Unsupervised Outlier Detection

Alternative ways to tackle the problem

1. To measure the performance by varing threshold
   - **AUC(Area Under the Curve) of ROC(Recieve Operating Characteristic) curve**
     - Measures FPR(False Positive Rate) and TPR(True Positive Rate) as the threshold varies
2. To set a specific threshold
   - **Extreme-value analysis**
     - Determines the statistical tails of the underlying distribution (i.e., set a threhold for the outlierness distribution in order to determine outliers)
     - And then, apply the binary classification measure (e.g., F_1-score) to evaluate the performance

Evaluatoin of Unsupervised Outlier Detection

Example of extreme-value analysis

- One of the simplest extreme-value analysis methods is to use quartiles proposed by Tukey et al. (1977) which does not requires distribution assumptions
- Tukey et al. (1977) classifies the observation as an outlier when the value of the observation is out of the following range

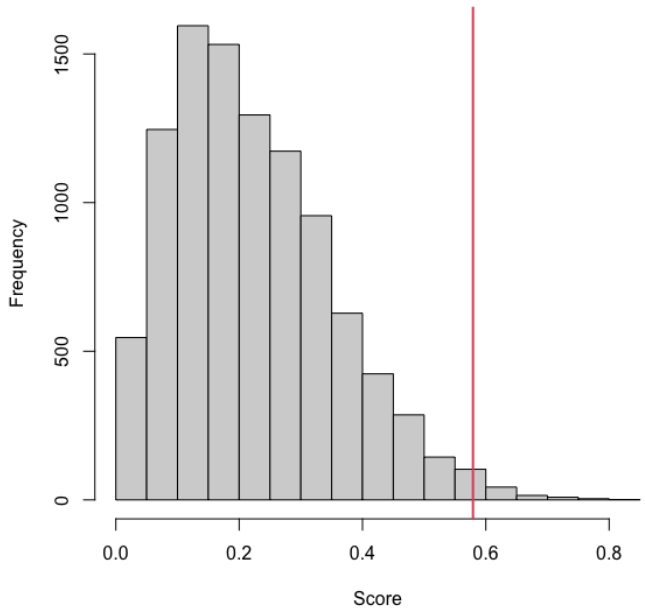
$$[Q_1 - k(Q_3 - Q_1),\ Q_3 + k(Q_3 - Q_1)]$$

  - Where, $Q_t, t \leq 3$ are the first, second, and third quartiles respectively
  - And $k = 1.5$ is used in general

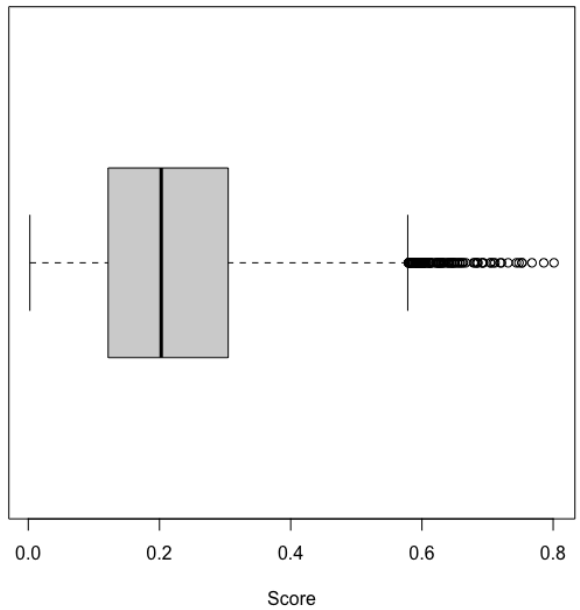
- However, when this is applied to degree of outlierness, only observations with values greater than the right range are considered as outliers

Evaluatoin of Unsupervised Outlier Detection

- Below figures show the application of Tukey's method on the right-skewed distribution
- It is based on the assumption that outliers will have much greater outlier scores than normal observations



(a) Histogram of outlier score and its threshold by applying Tukey's method



(b) Boxplot of outlier score and its threshold by applying Tukey's method

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Description of experiments

- In this section, the performance of subspace outlier detection techniques applied to the real-world datasets is provided compared to, the representative classical outlier detection method, LOF
- The data used in the analysis are the datasets suggested by Campos et al. (2016) for the evaluation of outlier detection that include outlier labels
- For evaluation, the F_1-score calculated based on the threshold obtained by applying Tukey et al. (1977)'s method to the degree of outlierness for each technique is used.

Evaluation measure

- F_1-score is the harmonic mean of the precision and recall
- $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$
  - precision $= \frac{\text{tp}}{\text{tp} + \text{fp}}$
  - recall $= \frac{\text{tp}}{\text{tp} + \text{fn}}$
- Since it only accounts for the performance of positive class, it is mainly used for the evaluation that needs an accurate prediction on a specific class

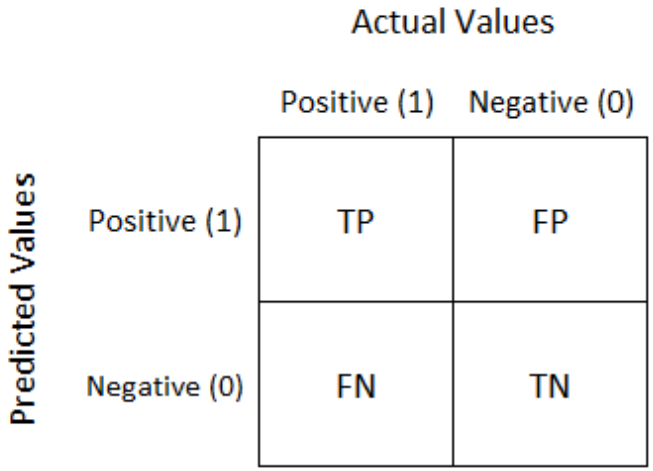


Figure: Confusion Matrix

Real-world Data Analysis

Table: Datasets used in the experiments

Dataset	Dimension	Sample size	Number of outliers	Percentage of outliers
Glass	7	214	9	4.21
Pima	8	526	26	4.94
WBC	9	454	10	2.20
Lymphography	19	148	6	4.05
Ionosphere	32	351	126	35.90
WPBC	33	198	47	23.74
SpamBase	57	2934	146	4.98
Arrhythmia	259	256	12	4.69

Real-world Data Analysis

Table: F_1-score results

Dataset	LOF	FB	IF	SOD	OUTRES	OutRank	HiCS
Glass	**0.2500**	0.1333	0.1429	0.2381	0.1395	0	0.1538
Pima	**0.1800**	0.0741	0.1176	0.0833	0	0	0.0952
WBC	0.3721	0.1739	0.3000	**0.4167**	0.3462	0.0441	0.3721
Lymphography	0.5882	0.2500	**0.6154**	0.5000	none	0.0714	0.2000
Ionosphere	0.2739	0.1870	0.3007	0.3087	none	0.3422	**0.4285**
WPBC	0.0615	0.0656	0	0.1538	none	**0.2078**	0.1667
SpamBase	0.1184	0	**0.3353**	0.1036	none	0.3115	0.1960
Arrhythmia	0.1764	0.2609	0.2400	0.2500	none	0.1099	**0.3200**

In each row, the highest value is shown in bold.

If a methodology takes more than 24 hours to apply, it is written as 'none'

Real-world Data Analysis

Table: User-defined elements used in the experiments

Dataset	LOF	FB	IF	SOD			OUTRES		OutRank	HiCS		
	k	m	m	k	l	α	ϵ	α	SC	α	m	t
Glass	21	20	20	21	15	1.1	15	0.1	DOC	0.1	50	32
Pima	53	20	20	53	37	1.1	15	0.1	DOC	0.1	50	32
WBC	45	20	20	45	32	1.1	15	0.1	DOC	0.1	50	32
Lymphography	15	20	20	15	10	1.1	15	0.1	DOC	0.1	50	32
Ionosphere	35	40	40	35	25	1.1	15	0.1	DOC	0.1	50	100
WPBC	20	40	40	20	14	1.1	15	0.1	DOC	0.1	50	100
SpamBase	293	100	100	293	205	1.1	15	0.1	DOC	0.1	50	100
Arrhythmia	26	200	200	26	18	1.1	15	0.1	DOC	0.1	50	200

Table of Contents

1. Introduction
   - Rapid Introduction to Outlier Detection
   - Concentration Effect: A Motivation of High-dimensional Outlier Detection
2. Subspace Outlier Detection
   - Random Subspace Search
   - Local Subspace Search
   - Global Subspace Search
3. Application and Evaluation of Subspace Outlier Detection
   - Evlauation of Unsupervised Outlier Detection
   - Real-world Data Anlaysis
4. Conclusion

Conclusion

- In this presentation, an overview of the classical outlier detection method was introduced and the motivation of high-dimensional outlier detection was investigated in the perspective of concentration effect
- Subspace outlier detection was an approach that tries to overcome the problem of high-dimensional data for outlier detection by selecting specific subspaces
- 7 representative techniques of subspace outlier detection were categorized into 3 types: Random subspace search, Local subspace search, Global subspace search
- And all method was summarized according to the specific criteria which are the way of subspace search, the result of subspace selection, and the outlierness measurement
- In the end, the evaluation of unsupervised outlier detection method and real-world data analysis were provided

Q&A

References I

- Basu, Sabyasachi and Martin Meckesheimer (2007). "Automatic outlier detection for time series: an application to sensor data". In: *Knowledge and Information Systems* 11(2), pp. 137–154.
- Beyer, Kevin et al. (1999). "When is "nearest neighbor" meaningful?" In: *International conference on database theory*. Springer, pp. 217–235.
- Breunig, Markus M et al. (2000). "LOF: identifying density-based local outliers". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104.
- Campos, Guilherme O et al. (2016). "On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study". In: *Data mining and knowledge discovery* 30(4), pp. 891–927.
- Chandola, Varun, Arindam Banerjee, and Vipin Kumar (2009). "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41(3), pp. 1–58.

References II

- Durrant, Robert J and Ata Kabán (2009). "When is 'nearest neighbour'meaningful: A converse theorem and implications". In: *Journal of Complexity* 25(4), pp. 385–397.
- Eskin, Eleazar et al. (2002). "A geometric framework for unsupervised anomaly detection". In: *Applications of data mining in computer security*. Springer, pp. 77–101.
- Fawcett, Tom and Foster Provost (1997). "Adaptive fraud detection". In: *Data mining and knowledge discovery* 1(3), pp. 291–316.
- Hawkins, Douglas M (1980). *Identification of outliers*. Vol. 11. Springer.
- Keller, Fabian, Emmanuel Muller, and Klemens Bohm (2012). "HiCS: High contrast subspaces for density-based outlier ranking". In: *2012 IEEE 28th international conference on data engineering*. IEEE, pp. 1037–1048.
- Kriegel, Hans-Peter et al. (2009). "Outlier detection in axis-parallel subspaces of high dimensional data". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, pp. 831–838.

References III

- Lazarevic, Aleksandar and Vipin Kumar (2005). "Feature bagging for outlier detection". In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 157–166.
- Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou (2008). "Isolation forest". In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE, pp. 413–422.
- Müller, Emmanuel, Matthias Schiffer, and Thomas Seidl (2011). "Statistical selection of relevant subspace projections for outlier ranking". In: *2011 IEEE 27th international conference on data engineering*. IEEE, pp. 434–445.
- Müller, Emmanuel et al. (2012). "Outlier ranking via subspace analysis in multiple views of the data". In: *2012 IEEE 12th international conference on data mining*. IEEE, pp. 529–538.

References IV

- Nguyen, Hoang Vu et al. (2013). "CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection". In: *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, pp. 198–206.
- Penny, Kay I and Ian T Jolliffe (2001). "A comparison of multivariate outlier detection methods for clinical laboratory safety data". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 50(3), pp. 295–307.
- Tukey, John W et al. (1977). *Exploratory data analysis*. Vol. 2. Reading, Mass.
- Zimek, Arthur, Erich Schubert, and Hans-Peter Kriegel (2012). "A survey on unsupervised outlier detection in high-dimensional numerical data". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 5(5), pp. 363–387.