Shapley Values of Reconstruction Errors of PCA for Explaining Anomaly Detection

Naoya Takeishi (RIKEN AIP)

8 November 2019
Workshop on Learning and Mining with Industrial Data, Beijing

Preprint available at arxiv.org/abs/1909.03495

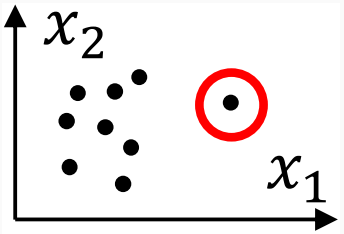
Background: Anomaly detection and localization

Anomaly detection

Anomaly detection is a fundamental problem of machine learning for industrial data, with many applications such as fault detection, intrusion detection, etc.

Problem: Anomaly detection (informal)

To find unexpected behavior from data.



Methodologies for anomaly detection (see, e.g., [Chandola+ 09])

- Rule-/model-based (limit-check, logical rules, physical models, etc.)
- Density-based (nearest neighbor, local outlier factor, etc.)
- One-class classification (OCSVM, etc.)
- Subspace-based (PCA, autoencoders, etc.)
  - easy-to-apply, works well for correlated multidimensional data

A practice in subspace-based anomaly detection

First, train an encoder-decoder model (PCA, autoencoders, etc.) using normal data as training data:

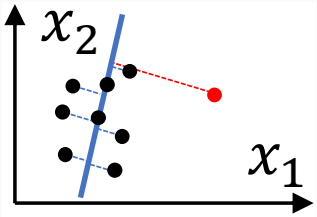
$$\underset{\text{original signal}}{x} \longrightarrow_{\text{encoder } f} \underset{\text{latent representation}}{z = f(x)} \longrightarrow_{\text{decoder } g} \underset{\text{reconstructed signal}}{\tilde{x} = g(z)}$$

If x is normal, x will be reconstructed well ($\tilde{x} \approx x$) also on test examples. Otherwise (i.e., x anomalous), the reconstruction error will be large.

$$\text{(reconstruction error)} = \|\tilde{x} - x\|$$

Simplest practice: Principal component analysis (PCA)

1. Train a PCA model on normal data.
2. Watch reconstruction errors on test examples.
3. Large reconstruction errors imply anomalies.

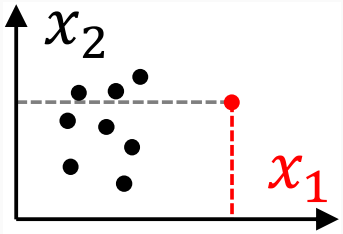


Anomaly localization

In practice, we want not only to detect, but also *to localize* anomalies.

Problem: Anomaly localization (informal)

To find (the most) anomalous features.



In subspace-based methods, the simplest way for localization is to watch each component of reconstruction errors. For d-feature data $\boldsymbol{x} \in \mathbb{R}^d$,

$$(\text{reconstruction error}) = \|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2 = \sqrt{(\tilde{x}_1 - x_1)^2 + \cdots + (\tilde{x}_d - x_d)^2}$$

$$\boxed{(\text{anomalous feature}) = \arg\max_i \ (\tilde{x}_i - x_i)^2}$$

However, the feature with largest reconstruction error is not necessarily anomalous. Perhaps, it was not reconstructed well *only occasionally* ☹

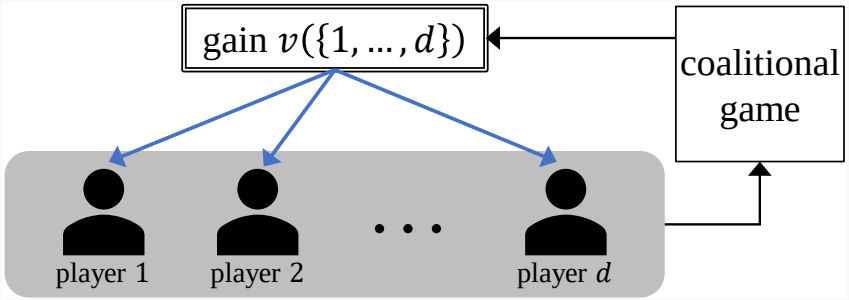
$\rightarrow$ Need a better way to localize anomalies using reconstruction errors.

Proposed method: Shapley values of reconstruction errors

Review: Shapley value

Shapley value [Shapley 53]

A (somewhat good) way to distribute the total gain of a coalitional game to its players.



Suppose there are d players, and let v : subset of $\{1,\ldots,d\} \to \mathbb{R}$ be the gain of game (e.g., $v(\{1,\ldots,d\})$ is for when everyone participated in).

The Shapley value of the i-th player (under gain function v) is given as the averaged effect for the i-th player to participate in the game, i.e.,

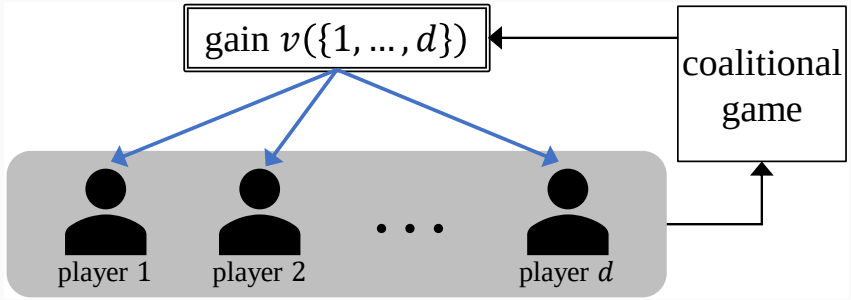
$$\varphi_i(v) = \sum_{S \subseteq \{1,\ldots,d\} \setminus \{i\}} \binom{d-1}{|S|}^{-1} \Big(v(S \cup \{i\}) - v(S) \Big)$$

It has been used for explaining ML [Štrumbelj&Kononenko 10,14; Lundberg&Lee 17].

Idea: Shapley value of reconstruction errors

Shapley value [Shapley 53]

A (somewhat good) way to distribute the total gain of a coalitional game to its players.

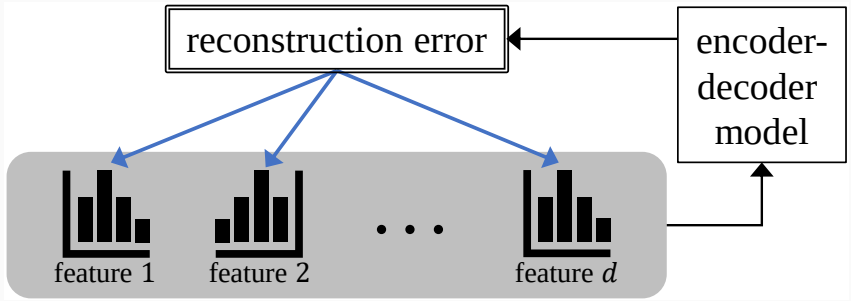


Which *player* contributed to the gain?

↓

Our idea: Shapley errors

To compute the Shapley value of reconstruction errors for anomaly localization.



Which *feature* contributed to the reconstruction error?

Challenge 1: How to define the gain function?

Shapley value for gain function v (again):

$$\varphi_i(v) = \sum_{S \subseteq \{1,\dots,d\} \setminus \{i\}} \binom{d-1}{|S|}^{-1} \Big(v(S \cup \{i\}) - v(S) \Big)$$

In our case (for reconstruction errors), how $v(\cdot)$ should be defined?

$\rightarrow$ Define v by partially-marginalized reconstruction errors (similarly to previous studies [Štrumbelj&Kononenko 10,14; Lundberg&Lee 17]).

$$v(S) = \mathbb{E}_{p(\boldsymbol{x}_{S^c} | \boldsymbol{x}_S)} \left[\| \tilde{\boldsymbol{x}} - \boldsymbol{x} \|_2^2 \right]$$

S^c	complement of S
$\boldsymbol{x}_{S^c}$	subvector of $\boldsymbol{x}$, indices corresponding to the elements of S^c
	e.g., $d = 3$, $S = \{1, 3\} \Rightarrow S^c = \{2\}$, $\boldsymbol{x}_S = [x_1, x_3]^\top$, $\boldsymbol{x}_{S^c} = [x_2]$

Challenge 2: Dependency of features

The gain function for reconstruction errors:

$$v(S) = \mathbb{E}_{p(\boldsymbol{x}_{S^c}|\boldsymbol{x}_S)}\left[\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2\right]$$

Can we compute $\mathbb{E}_{p(\boldsymbol{x}_{S^c}|\boldsymbol{x}_S)}[\cdot]$?

→ Usually, features are assumed to be independent [Štrumbelj&Kononenko 14; Ribeiro+ 16; Lundberg&Lee 17], which is inappropriate in our case.

→ Focus on PCA: $p(\boldsymbol{x}_{S^c} \mid \boldsymbol{x}_S)$ becomes Gaussian [Tipping&Bishop 99].

$$p(\boldsymbol{x}_{S^c} \mid \boldsymbol{x}_S) = \mathcal{N}_{\boldsymbol{x}_{S^c}}\left(\boldsymbol{C}_{S^c,S}\boldsymbol{C}_S^{-1}\boldsymbol{x}_S,\ \boldsymbol{C}_{S^c} - \boldsymbol{C}_{S^c,S}\boldsymbol{C}_S^{-1}\boldsymbol{C}_{S^c,S}^\top\right)$$

$\boldsymbol{C}_S,\ \boldsymbol{C}_{S^c}$	submatrices of $\boldsymbol{C} = \sigma^2\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^\top$
$\boldsymbol{W}$	factor-loading matrix of PCA
σ^2	observation noise variance of PCA

Shapley value of PCA's reconstruction errors

In a nutshell, we compute

$$\varphi_i(v) = \sum_{S \subseteq \{1,\dots,d\} \setminus \{i\}} \binom{d-1}{|S|}^{-1} \Big(v(S \cup \{i\}) - v(S) \Big),$$

where (the definitions of $\boldsymbol{B}$, $\boldsymbol{V}$, and $\boldsymbol{m}$ are omitted here)

$$\begin{aligned} v(S) &= \mathbb{E}_{p(\boldsymbol{x}_{S^c} | \boldsymbol{x}_S)} \left[\| \tilde{\boldsymbol{x}} - \boldsymbol{x} \|_2^2 \right] \\ &= \operatorname{trace}\left((\boldsymbol{I} - \boldsymbol{B}_{S^c}) \boldsymbol{V}_{S^c} \right) + \operatorname{trace}\left((\boldsymbol{I} - \boldsymbol{B}_{S^c}) \boldsymbol{m}_{S^c} \boldsymbol{m}_{S^c}^\top \right) \\ &\quad - 2 \operatorname{trace}(\boldsymbol{B}_{S^c,S} \boldsymbol{x}_S \boldsymbol{m}_{S^c}^\top) + \operatorname{trace}\left((\boldsymbol{I} - \boldsymbol{B}_S) \boldsymbol{x}_S \boldsymbol{x}_S^\top \right), \end{aligned}$$

and the summation over subsets is approximated by Monte Carlo method.

Finally, an anomalous feature is determined by $\boxed{\arg\max_i \ \varphi_i(v)}$.

Preliminary experiments

Performance on synthetic dataset: Setting

Verified localization performance on synthetic anomalies.

Baseline (anomalous feature) $= \arg\max_i \ |\tilde{x}_i - x_i|$

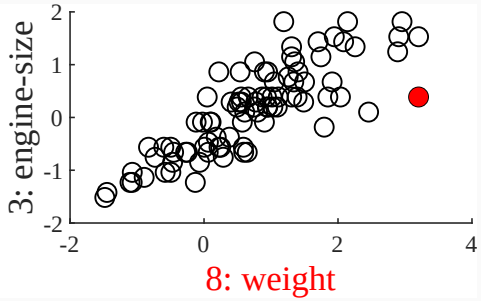
Proposed (anomalous feature) $= \arg\max_i \ \varphi_i(v)$

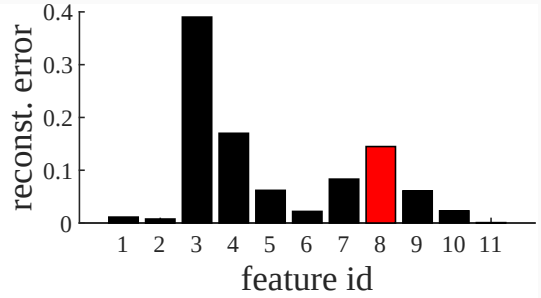
Dataset 2004 New Car and Truck Data (JSE Data Archive)
$n = 428$ observations, $d = 11$ features w/o missing values

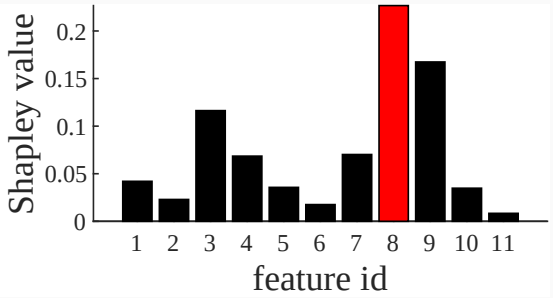
01: price	02: cost	03: engine-size	04: #cylinders
05: horsepower	06: city-mpg	07: highway-mpg	08: weight
09: wheel-base	10: length	11: width	

Inserted artificial anomalies by flipping the value of a feature to its max/min value, for $j = 1, \ldots, 428$ and $i = 1, \ldots, 11$ at each trial.

Performance on synthetic dataset: Results (1)







Example: Anomaly was inserted to $i = 8$ of a datapoint. Reconstruction error (center) fails localize it, but its Shapley value (right) succeeds to localize.

Performance on synthetic dataset: Results (2)

Hits@k (the rate that anomalous feature is correctly localized by looking at the top-k values) for the two experimental cases over many trials.

	flip w/ max		flip w/ min	
	Hits@1	Hits@3	Hits@1	Hits@3
reconstruction error	.316	.605	.271	.471
Shapley value	**.484**	**.801**	**.484**	**.710**

Behavior on real-world datasets

Investigated correlation between reconstruction error and Shapley value.

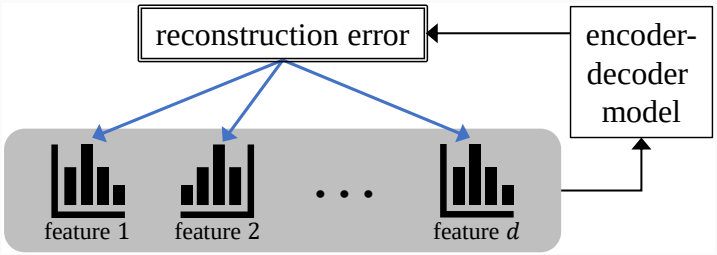
Dataset Outlier Detection Datasets (OODS) `odds.cs.stonybrook.edu`
Picked up the ones on which PCA-based detection worked.

Results In some cases, the correlation is not strong, which suggests that both values should be watched.

dataset			correlations		
name	d	n	r_{all}	r_{normal}	$r_{\text{anomalous}}$
Cardio	21	1831	.866	.893	.797
ForestCover	10	286048	.756	.536	.808
Ionosphere	33	351	.984	.986	.985
Mammography	6	11183	.854	.268	.854
Musk	166	3062	.945	.987	.949
Satimage-2	36	5803	.975	.993	.981
Shuttle	9	49097	.869	.958	.893
Vowels	12	1456	.883	.833	.877
WBC	30	278	.956	.955	.943
Wine	13	129	.817	.785	.657

Summary

Anomaly localization by Shapley values of reconstruction errors



Problem Anomaly localization — which feature is anomalous?

Idea Watch the Shapley value of reconstruction errors.

Challenge Features are usually dependent.

Proposal Focus on PCA, for which the feature dependence is Gaussian and the gain for the Shapley value can be computed exactly.

Future work Extension for non-linear, non-Gaussian cases (e.g., VAEs).
Why reconstruction error fails to localize?
More efficient computation. etc.

Preprint available at `arxiv.org/abs/1909.03495`

Appendix

Detailed calculation of the Shapley value for PCA

$$\varphi_i(v) = \frac{1}{d!} \sum_{O \in \pi(1,\ldots,d)} \Big(v(Pre_i(O) \cup \{i\}) - v(Pre_i(O)) \Big),$$

$\pi(1,\ldots,d)$ is the set of permutations of $(1,\ldots,d)$. $Pre_i(O)$ denotes the set of feature indices that precede i in order O.

The summation is approximated by the Monte Carlo method.

$$\begin{aligned} v(S) &= \mathbb{E}_{p(\boldsymbol{x}_{S^c}|\boldsymbol{x}_S)} \left[\|\tilde{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 \right] \\ &= \operatorname{trace}\left((\boldsymbol{I} - \boldsymbol{B}_{S^c}) \boldsymbol{V}_{S^c} \right) + \operatorname{trace}\left((\boldsymbol{I} - \boldsymbol{B}_{S^c}) \boldsymbol{m}_{S^c} \boldsymbol{m}_{S^c}^\top \right) \\ &\quad - 2 \operatorname{trace}(\boldsymbol{B}_{S^c,S} \boldsymbol{x}_S \boldsymbol{m}_{S^c}^\top) + \operatorname{trace}\left((\boldsymbol{I} - \boldsymbol{B}_S) \boldsymbol{x}_S \boldsymbol{x}_S^\top \right), \end{aligned}$$

$$\boldsymbol{C} = \sigma^2 \boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^\top, \quad \boldsymbol{B} = \boldsymbol{W}(\boldsymbol{W}^\top \boldsymbol{W})^{-1} \boldsymbol{W}^\top,$$

$$\boldsymbol{m}_{S^c} = \boldsymbol{C}_{S^c,S} \boldsymbol{C}_S^{-1} \boldsymbol{x}_S, \quad \boldsymbol{V}_{S^c} = \boldsymbol{C}_{S^c} - \boldsymbol{C}_{S^c,S} \boldsymbol{C}_S^{-1} \boldsymbol{C}_{S^c,S}^\top.$$

$\boldsymbol{W} \in \mathbb{R}^{d \times p}$ is the factor-loading matrix of PCA, σ^2 is the observation noise variance. $\cdot_S$ denotes the submatrix/subvector corresponding to the elements of $S \subseteq \{1,\ldots,d\}$. S^c is the complement of S.

Additional results