LoPAD: A **Lo**cal **P**rediction Approach to **A**nomaly **D**etection

Sha Lu, Lin Liu, Jiuyong Li, Thuc Duy Le, and Jixue Liu

University of South Australia

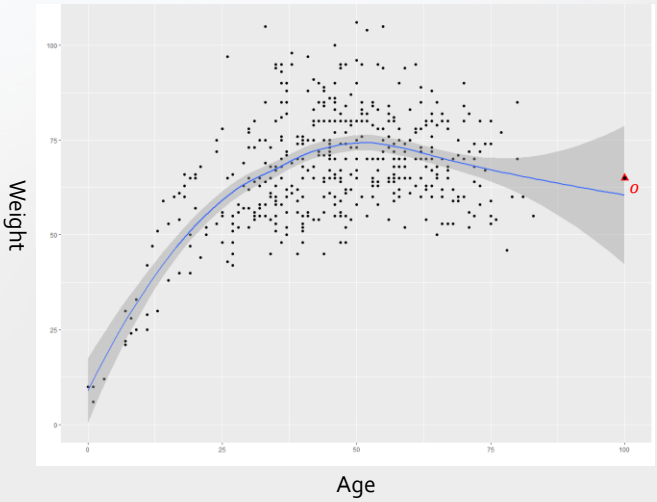
1

Background

- Anomalies are *patterns in data that do not conform to a well-defined notion of normal behavior*.
- Anomaly detection methods highly rely on *the assumption of anomalies*.
- Proximity-based anomaly detection
- Dependency-based anomaly detection

2

Proximity-based and Dependency-based Approaches



Q: what is the detection results of *o* by the two types of methods?

Based on different assumptions,

- Proximity-based: Yes ✓
- Dependency-based: No ✗

3

The LoPAD Framework

- value-wise deviation evaluates how well an object follows the dependency around a specific variable

Definition 1. *(Value-wise Deviation) Given an object x_i, its value-wise deviation with respect to variable X_j is defined as:*

$$\delta_{ij} = |x_{ij} - \widehat{x}_{ij}| \quad (1)$$

where x_{ij} is the observed value of X_j in x_i, and

$$\widehat{x}_{ij} = g(\boldsymbol{X}' = \boldsymbol{x}'_i) \quad (2)$$

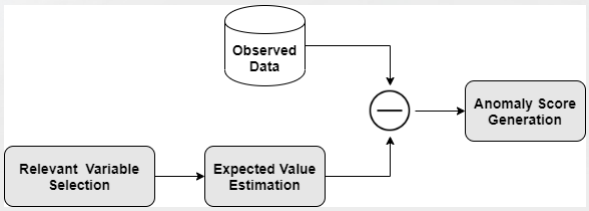
is the expected value of X_j estimated using the function $g()$ based on the values on other variables $\boldsymbol{X}' \subseteq \boldsymbol{X} \setminus \{X_j\}$.

- vector-wise deviation evaluates how an object collectively follows the dependencies

Definition 2. *(Vector-wise Deviation) The vector-wise deviation of object x_i is the aggregation of all its value-wise deviations calculated using a combination function as follows:*

$$\delta_i = combine(\delta_{i1}, \ldots, \delta_{im}) \quad (3)$$

Definition 3. *(Problem Definition) Given a dataset $\boldsymbol{D}$ with n objects and a user specified parameter k, our goal is to detect the top-k ranked objects according to the descending order of vector-wise deviations as anomalies.*



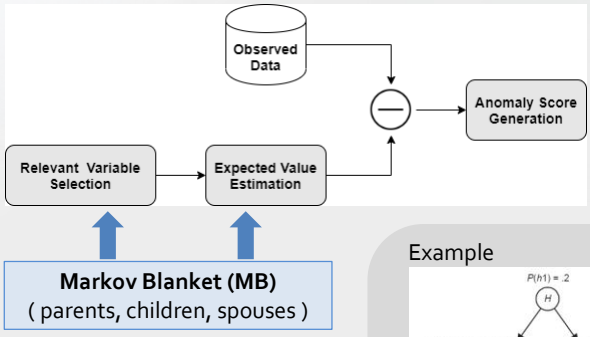
1. **Relevant Variable Selection**: to select the optimal relevant variables for a target variable.
2. **Expected Value Estimation**: to estimate the expected value of a variable in an object using the selected variables.
3. **Anomaly Score Generation**: to obtain vector-wise deviation, i.e., anomaly score by applying a combination function over value-wise deviations.

4

The LoPAD Framework

Two Challenges:

1. How to capture the complete dependency and also keep the most relevant variables at the same time?
2. How to select relevant variables to assure an accurate prediction of expected value?



Markov Condition:
In a BN, given all its parents, a node X is conditionally independent of all its non-descendant nodes.

Example

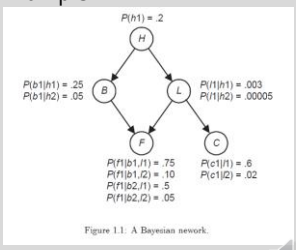


Figure 1.1: A Bayesian nework.

$MB(B) = \{H, F, L\}$

5

The LoPAD Algorithm

Algorithm 1: The LoPAD Algorithm

Input: D, a dataset with n objects and a set of m variables, denoted as $\mathbf{X}$; k, the number of anomalies to output
Output: top k detected anomalies

Relevant Variable Selection
1: initialize deviation matrix $\boldsymbol{\Delta}_{n \times m}$
2: **for** each $X_j \in \mathbf{X}, j \in \{1, \ldots, m\}$ **do**
3: discovery $MB(X_j)$ using fast-IAMB algorithm *//relevant variable selection*

Expected Value Estimation
4: train a prediction (CART) model g_j: $X_j = g_j(MB(X_j))$
5: **for** each $\mathbf{x}_i \in \mathbf{D}, i \in \{1, \ldots, n\}$ **do**
6: predict $\widehat{x}_{ij}$ with g_j using Equation 2
7: compute δ'_{ij} using Equation 1 *//value-wise deviation*
8: **end for**
9: **end for**

Anomaly Score Generation
10: normalize $\boldsymbol{\Delta}$ *//normalization*
11: **for** each $\mathbf{x}_i \in \mathbf{D}, i \in \{1, \ldots, n\}$ **do**
12: compute anomly score δ_i using Equation 6 *//vector-wise deviation*
13: **end for**
14: output top-k scored objects based on descending order of $\delta_i (i \in \{1, \ldots, n\})$

6

Experimental Results

Table 1: The Summary of 4 Synthetic and 13 Real-world Datasets

	MAGIC -NIAB	ECILI -70	MAGIC -IRRI	ARTH -150	breast cancer	wine	biode-gradation	bank	spam-base	AID362	back-door	cal-Tech16	census	secom	arrhy -thmia	mnist	ads
#sample	5000	5000	5000	5000	448	4898	359	4040	2815	4279	56560	806	45155	1478	343	1038	2848
#variable	44	46	64	107	9	11	41	51	57	144	190	253	409	590	680	784	1446

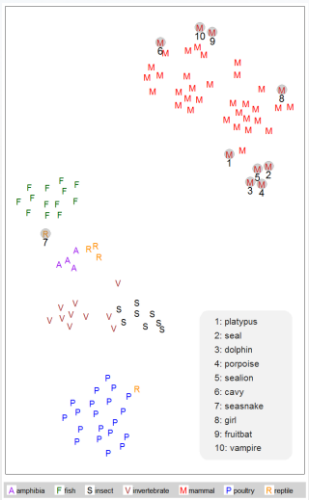
- **Comparison methods**
  1. Dependency-based: ALSO, COMBN
  2. Proximity-based: MBOM, iForest, LOF
- Evaluation measure: ROC AUC
- LoPAD improves ROC AUC of comparison methods from 4.2% to 14.6%
- The p-values of Wilcoxon rank sum test are below 0.05 except COMBN
- The average size of MBs are much smaller than the original dimensions.

Table 2: Experimental Results (ROC AUC)

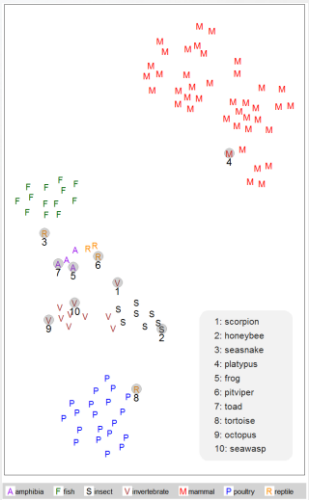
Dataset	average size of MBs	LoPAD	ALSO	MBOM	COMBN	iForest	LOF
MAGIC-NIAB	8.0	**0.826±0.033**	0.775±0.106	0.817±0.052	0.719±0.099	0.780±0.035	0.819±0.028
ECOLI70	6.5	0.987±0.013	**0.994±0.008**	0.992±0.008	0.988±0.013	0.799±0.027	0.972±0.014
MAGIC-IRRI	8.1	**0.917±0.051**	0.861±0.123	0.899±0.041	0.876±0.079	0.817±0.037	0.891±0.029
ARTH150	7.9	**0.986±0.011**	**0.986±0.017**	0.959±0.022	0.984±0.011	0.853±0.028	0.962±0.009
breast cancer	3.5	**0.996±0.004**	0.984±0.011	0.961±0.013	0.989±0.006	0.991±0.005	0.891±0.031
wine	8.9	**0.812**	0.782	0.800	0.722	0.754	0.782
biodegradation	14.8	**0.883±0.063**	0.855±0.084	0.808±0.105	0.856±0.082	**0.883±0.069**	0.868±0.083
bank	17.7	**0.750±0.038**	0.682±0.045	0.661±0.043	0.706±0.051	0.679±0.048	0.566±0.043
spambase	10.0	**0.821±0.038**	0.653±0.045	0.718±0.034	0.808±0.053	0.773±0.041	0.801±0.03
AID362	51.9	0.604	0.594	0.550	**0.674**	0.634	0.570
backdoor	92.4	**0.941±0.005**	0.922±0.009	0.765±0.027	-	0.794±0.035	0.748±0.018
calTech16	48.8	0.98 ±0.006	0.979±0.006	0.766±0.039	0.981±0.006	**0.983±0.004**	0.491±0.086
census	69.3	**0.663±0.011**	0.642±0.012	0.608±0.013	-	0.575±0.02	0.502±0.013
secom	35	0.596±0.067	0.594±0.074	0.551±0.066	**0.610±0.081**	0.533±0.074	0.538±0.086
arrhythmia	61.7	**0.914**	0.892	0.563	-	0.844	0.906
mnist	65.3	**0.997±0.002**	0.991±0.004	0.606±0.099	-	0.996±0.003	0.958±0.044
ads	68.7	**0.932±0.032**	0.894±0.032	0.864±0.033	-	0.754±0.06	0.851±0.036
Average AUC		**0.859±0.027**	0.828±0.041	0.758±0.043	0.826±0.048	0.791±0.035	0.772±0.039
AUC Improvement		-	4.2%	15.3%	2.6%	9.1%	14.6%
Wilcoxon rank sum test p-value		-	0.0005	0.0001	0.0599	0.0007	0.0002

7

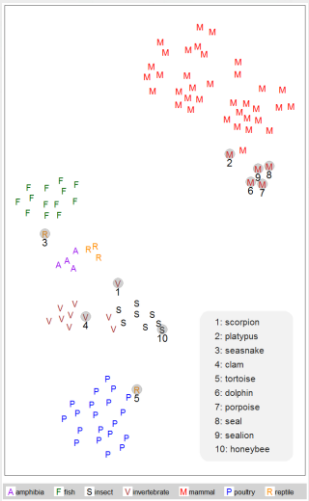
A Case Study: zoo dataset



LOF



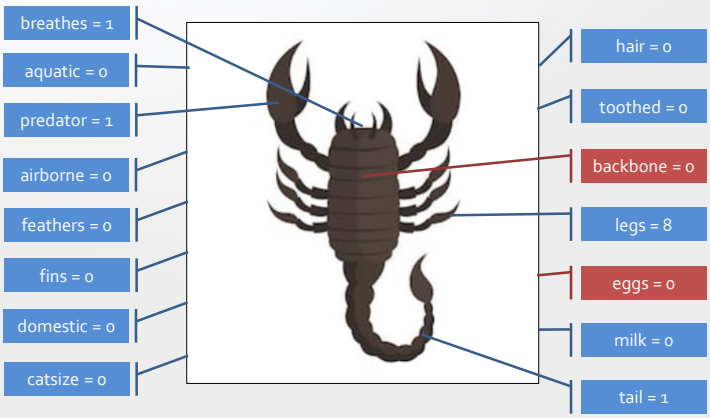
Weighted kNN



LoPAD

8

The Interpretation of Detected Anomalies



TOP 2 value-wise deviations Of scorpion

1
- **normal**: if tail = 1, then backbone = 1
- **scorpion**: tail = 1, backbone = 0
- Only 1 in the dataset

2
- **normal**: if milk = 0 , then eggs = 1
- **scorpion**: milk = 0, eggs = 0
- 1 out of 2 in the dataset

Totally, there are 16 variables, and one of them (venomous) is irrelevant.

9

Contributions

- Through *introducing Markov Blanket* into dependency-based anomaly detection, we decompose the high dimensional unsupervised anomaly detection problem into local feature selection and prediction problems, which also provide better interpretation of detected anomalies.
- We develop *an anomaly detection framework, LoPAD*, to efficiently and effectively discover anomalies in high dimensional data of different types.
- We present *an instantiated algorithm* based on the LoPAD framework and conduct extensive experiments on a range of synthetic and real-world datasets to demonstrate the effectiveness and efficiency of LoPAD.

10

Questions?

11

Running Time

Table 3: Average Running Time(in seconds)

Dataset	LoPAD	ALSO	MBOM	COMBN	iForest	LOF	Dataset	LoPAD	ALSO	MBOM	COMBN	iForest	LOF
MAGIC-NIAB	12.8	35.5	28.4	2.5	1.2	1.7	ECOLI70	12.7	33.6	23.8	2.3	1.2	1.5
MAGIC-IRRI	14.7	61.4	41.4	5.5	1.5	2.0	ARTH150	20.0	164.8	68.9	10.0	2.1	2.7
breast cancer	0.7	1.3	0.3	0.01	0.37	0.04	wine	9.2	14.0	5.3	0.3	0.6	0.5
biodegradation	4.8	7.6	1.6	0.6	0.39	0.04	bank	14.3	23.4	18.6	7.9	1.0	0.8
spambase	11.9	24.5	13.9	3.2	0.8	0.4	AID362	116.6	123.3	160.3	591.9	2.1	1.7
backdoor	907.0	1148.8	1136.8	-	167.3	11.1	calTech16	53.4	52.7	54.9	558.0	0.8	0.1
census	2582.1	6041	4810.5	-	382.3	313.8	secom	75.1	454.8	133.8	1679.3	2.3	0.9
arrhythmia	375.6	150.0	370.9	-	0.84	0.09	mnist	366.4	267.4	389.1	-	1.9	0.4
ads	2265.0	2437.4	2486.3	-	53.7	4.7	**Average**	**402.5**	**649.5**	**573.2**	**238.4**	**36.5**	**20.1**

12

Sensitivity Tests

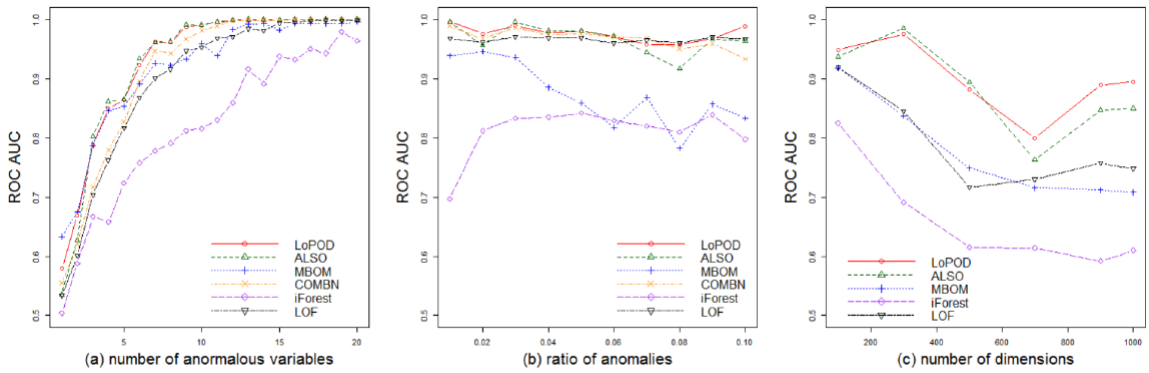


Fig. 1: The results of sensitivity experiments

13