Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection

Roy Kinamon and Eliran Elisha

Agenda

Introduction

Motivation

The Problem

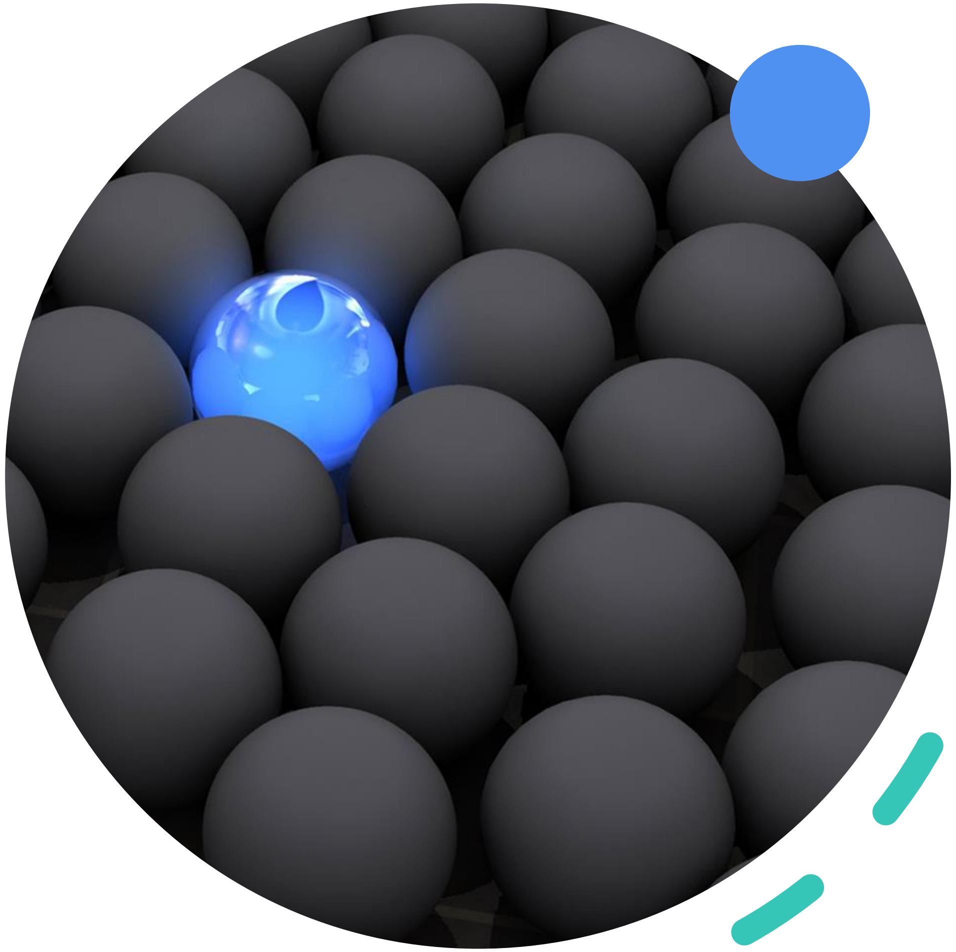
HBOS Paper

Suggested Improvement

Results

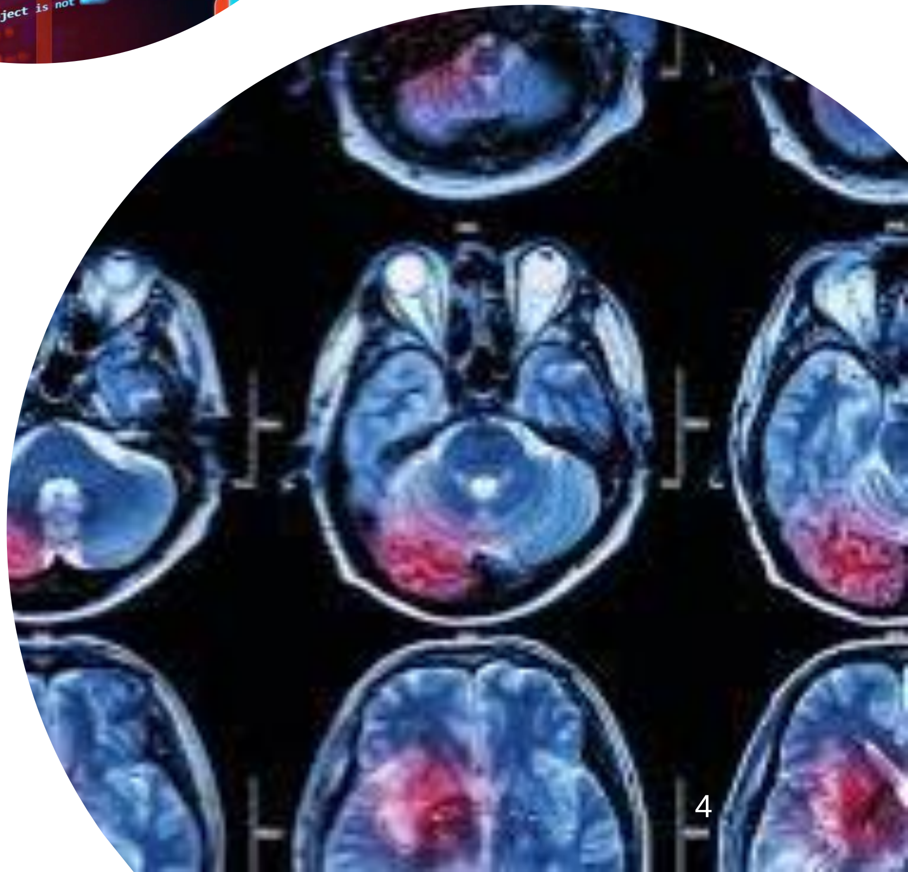
Introduction

Anomaly detection is the process of identifying data points that deviate from the expected patterns or behaviors of a given system or dataset. Anomalies are often referred to as outliers, novelties, or anomalies, and they can be caused by a variety of factors such as errors, fraud, unusual events, or changes in underlying trends.



Motivations

- The motivation behind anomaly detection is to detect unusual or suspicious events or behaviors that may indicate a potential problem or threat. It is an important task in many domains:
- finance - detect fraudulent transactions or identify abnormal trading behavior.
- Healthcare - identify rare diseases or unusual patient symptoms.
- Cybersecurity - detect network intrusions or identify unusual patterns of activity.
- industrial automation - can help identify equipment malfunctions or anomalies in manufacturing processes.



The Problem

Finding an effective solution for identifying anomalies in various datasets keepint

- Hige Precision in Anomaly detection
- Supporting unsupervised data
- Fast run time and low complexity

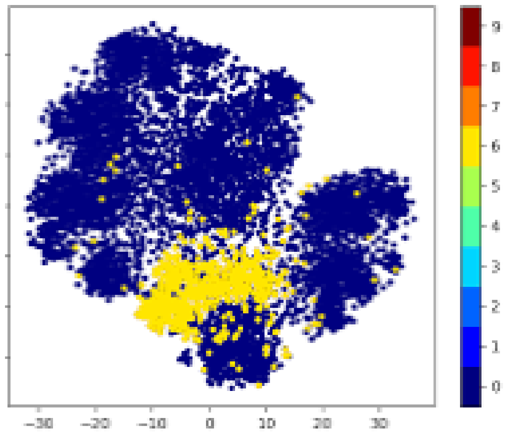
Histogram-based Outlier Score (HBOS)

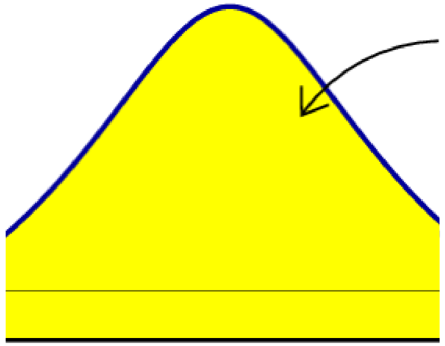
Markus Goldstein and Andreas Dengel

HBOS

- There are 3 main approaches to deal with anomaly detection problem:
  - distance-based algorithms like: KNN and Local Outlier Factor (LOF).
  - Clustering based algorithms like CBLOF and LDCOF are using k-means as a clustering algorithm.
  - Statistical methods: parametric and non-parametric models like GMM and kernel-density estimators (KDE).

While first 2 has high complexity and sensitivity to outliers, HBOS uses statistical methods





HBOS Calculation

- Bins with density estimation per feature
- Each histogram is normalized, max height=1
- Calculating HBOS using height of each Histogram

$$HBOS(p) = \sum_{i=0}^{d} \log\left(\frac{1}{hist_i(p)}\right)$$

- Linear computation time

HBOS Performance

HBOS outperforms or get close to best in half of the above datasets compared to all other algorithms

Alg.	b-cancer	pen-global	pen-local	letter	speech	satellite	thyroid	shuttle	aloi	kdd99
k-NN	<0.1	<0.1	2.4	0.3	5.7	2.0	2.6	106	166	538
k^{th}-NN	<0.1	<0.1	2.4	0.3	5.8	2.0	2.6	105	165	538
LOF	<0.1	<0.1	2.4	0.3	5.8	2.0	2.7	105	165	538
LOF-UB	<0.1	<0.1	2.6	0.3	5.9	2.1	2.8	107	167	539
COF	<0.1	0.1	2.8	0.5	9.0	2.5	3.1	107	169	539
INFLO	<0.1	<0.1	2.4	0.3	5.8	2.0	2.6	105	165	538
LoOP	<0.1	<0.1	2.5	0.3	5.8	2.0	2.6	105	165	538
LOCI	18	240	—	2572	25740	—	—	—	—	—
aLOCI	0.5	1.8	90	12.7	9.5	56	30	73	1137	298
CBLOF/LDCOF 10	<0.1	0.1	1.5	0.6	24.8	4.0	1.0	6.9	39.1	5.01
CBLOF/LDCOF 50	0.1	0.2	3.7	5.9	24.7	5.2	4.4	10.3	74.6	16.14
CMGOS-Red 10	0.5	0.2	1.7	1.1	82	4.6	1.2	7.0	40	5.15
CMGOS-Red 50	0.1	0.5	4.3	1.7	49	8.2	4.6	10.6	77	16.25
CMGOS-Reg 10	0.4	0.2	1.7	1.1	83	4.6	1.3	7.0	40	5.19
CMGOS-Reg 50	0.1	0.5	4.3	1.7	49	8.1	5.4	10.6	77	16.29
CMGOS-MCD 10	159	211	863	759	—	3821	1967	354	3003	491
CMGOS-MCD 50	735	519	1045	1441	—	4041	4159	1525	10933	8745
HBOS	<0.1	<0.1	<0.1	<0.1	0.5	<0.1	<0.1	<0.1	0.4	0.06
rPCA	<0.1	<0.1	0.2	<0.1	9.2	0.1	<0.1	0.3	1.5	21.8
oc-SVM	0.3	0.5	31	8.5	807	28	26	19639	59531	5480
η-oc-SVM	0.3	0.4	70	8.2	745	24	27	19087	58559	3310

doi:10.1371/journal.pone.0152173.t005

Real Time

Alg.	b-cancer	pen-global	pen-local	letter	speech	satellite	thyroid	shuttle	aloi	kdd99
k-NN	0.9791 ±0.0010	**0.9872** ±0.0055	0.9837 ±0.0018	0.8719 ±0.0176	0.4966 ±0.0101	**0.9701** ±0.0007	0.5956 ±0.0125	0.9424 ±0.0069	0.6502 ±0.0191	0.9747 ±0.0045
k^{th}-NN	0.9807 ±0.0008	0.9778 ±0.0142	0.9757 ±0.0069	0.8268 ±0.0228	0.4784 ±0.0007	0.9681 ±0.0015	0.5748 ±0.0128	0.9434 ±0.0101	0.6177 ±0.0189	**0.9796** ±0.0035
LOF	**0.9816** ±0.0024	0.8495 ±0.0679	**0.9877** ±0.0016	0.8673 ±0.0271	0.5038 ±0.0215	0.8147 ±0.1126	0.6470 ±0.0192	0.5127 ±0.0129	0.7563 ±0.0135	0.5964 ±0.0284
LOF-UB	0.9805 ±0.0020	0.8541 ±0.0777	0.9876 ±0.0013	0.9019 ±0.0030	0.5233 ±0.0134	0.8425 ±0.0839	0.6663 ±0.0103	0.5182 ±0.0124	0.7713 ±0.0045	0.5774 ±0.0159
COF	0.9518 ±0.0054	0.8695 ±0.1261	0.9513 ±0.0134	0.8336 ±0.0228	0.5218 ±0.0287	0.7491 ±0.0952	0.6505 ±0.0154	0.5257 ±0.0086	0.7857 ±0.0118	0.5548 ±0.0236
INFLO	0.9642 ±0.0171	0.7887 ±0.0540	0.9817 ±0.0024	0.8632 ±0.0250	0.5017 ±0.0191	0.8272 ±0.0761	0.6542 ±0.0158	0.4930 ±0.0175	0.7684 ±0.0142	0.5524 ±0.0222
LoOP	0.9725 ±0.0123	0.7684 ±0.0994	0.9851 ±0.0068	**0.9068** ±0.0078	**0.5347** ±0.0343	0.7681 ±0.0433	**0.6893** ±0.0149	0.5049 ±0.0035	**0.7899** ±0.0093	0.5749 ±0.0275
LOCI	0.9787	0.8877	—	0.7880	0.4979	—	—	—	—	—
aLOCI	0.8105 ±0.0883	0.6889 ±0.0345	0.8011 ±0.0615	0.6208 ±0.0220	0.4992 ±0.0348	0.8324 ±0.0372	0.6174 ±0.0221	**0.9474** ±0.0379	0.5855 ±0.0143	0.6552 ±0.0458

doi:10.1371/journal.pone.0152173.t002

Alg.	b-cancer	pen-global	pen-local	letter	speech	satellite	thyroid	shuttle	aloi	kdd99
CBLOF	0.2983 ±0.1492	0.3190 ±0.1155	0.6995 ±0.1407	0.6792 ±0.0386	0.5021 ±0.0680	0.5539 ±0.0692	0.5825 ±0.0384	0.9037 ±0.1263	0.5393 ±0.0154	0.6589 ±0.2098
uCBLOF	**0.9496** ±0.0390	**0.8721** ±0.0511	0.9555 ±0.0109	0.8192 ±0.0231	0.4692 ±0.0029	**0.9627** ±0.0038	0.5469 ±0.0212	**0.9716** ±0.0324	0.5575 ±0.0061	**0.9964** ±0.0016
LDCOF	0.7645 ±0.1653	0.5948 ±0.0825	0.9593 ±0.0145	0.8107 ±0.0244	0.4366 ±0.0099	0.9522 ±0.0325	0.5703 ±0.0232	0.8076 ±0.1814	0.5726 ±0.0146	0.9873 ±0.0034
CMGOS-Red	0.9140 ±0.0815	0.5693 ±0.1000	**0.9727** ±0.0141	0.7711 ±0.0614	0.5077 ±0.0158	0.9054 ±0.0233	0.4395 ±0.0402	0.5425 ±0.2446	0.5852 ±0.0161	0.7265 ±0.1027
CMGOS-Reg	0.8992 ±0.0643	0.6994 ±0.0681	0.9449 ±0.0510	**0.8902** ±0.0200	**0.5081** ±0.0161	0.9056 ±0.0233	0.6587 ±0.0268	0.5679 ±0.2402	**0.5855** ±0.0161	0.9797 ±0.0080
CMGOS-MCD	0.9196 ±0.0830	0.6265 ±0.0969	0.9038 ±0.0511	0.7848 ±0.0485	—	0.9120 ±0.0520	**0.8014** ±0.0436	0.6903 ±0.1670	0.5547 ±0.0160	0.9696 ±0.0416
Best NN	**0.9816** ±0.0024	**0.9872** ±0.0055	**0.9877** ±0.0016	**0.9068** ±0.0078	**0.5347** ±0.0343	**0.9701** ±0.0007	0.6893 ±0.0149	0.9474 ±0.0379	**0.7899** ±0.0093	0.9796 ±0.0035

doi:10.1371/journal.pone.0152173.t003

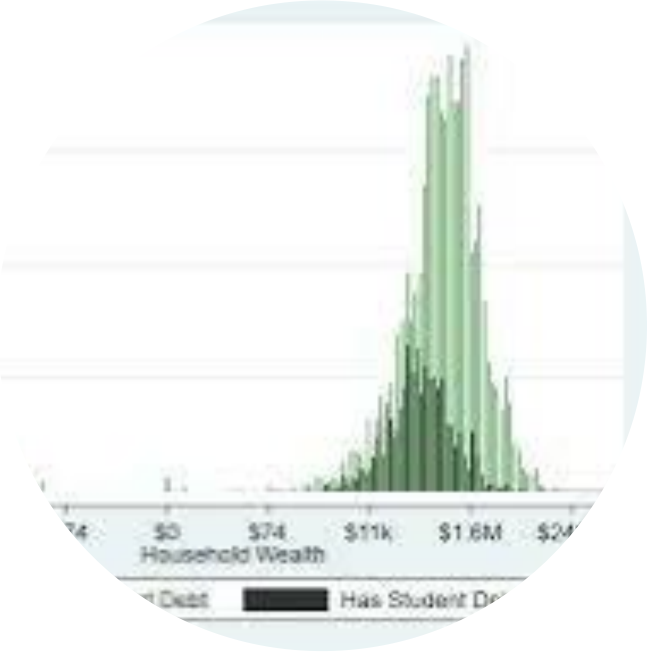
Alg.	b-cancer	pen-global	pen-local	letter	speech	satellite	thyroid	shuttle	aloi	kdd99
HBOS	**0.9827** ±0.0016	0.7477 ±0.0206	0.6798 ±0.0249	0.6216 ±0.0073	0.4708 ±0.0030	0.9135 ±0.0047	**0.9150** ±0.0123	**0.9925** ±0.0039	0.4757 ±0.0010	**0.9990** ±0.0007
rPCA	0.9664 ±0.0000	0.9375 ±0.0001	0.7841 ±0.0151	0.8095 ±0.0029	0.5024 ±0.0000	0.9461 ±0.0023	0.6574 ±0.0036	0.9963 ±0.0000	0.5621 ±0.0000	0.7371 ±0.0000
oc-SVM	0.9721 ±0.0102	0.9512 ±0.0436	0.9543 ±0.0130	0.5195 ±0.0382	0.4650 ±0.0021	0.9549 ±0.0021	0.5316 ±0.0152	0.9862 ±0.0002	0.5319 ±0.0021	0.9518 ±0.0050
η-oc-SVM	0.9581 ±0.0311	0.8993 ±0.0387	0.9236 ±0.0140	0.7298 ±0.1365	0.4649 ±0.0026	0.9430 ±0.0058	0.5625 ±0.0088	0.9848 ±0.0019	0.5221 ±0.0025	0.7945 ±0.0000
Best NN	0.9816 ±0.0024	**0.9872** ±0.0055	**0.9877** ±0.0016	**0.9068** ±0.0078	**0.5347** ±0.0343	**0.9701** ±0.0007	0.6893 ±0.0149	0.9474 ±0.0379	**0.7899** ±0.0093	0.9796 ±0.0035
Best Cluster	0.9496 ±0.0390	0.8721 ±0.0511	0.9727 ±0.0141	0.8902 ±0.0200	0.5081 ±0.0161	0.9627 ±0.0038	0.7843 ±0.0437	0.9716 ±0.0324	0.5855 ±0.0161	0.9964 ±0.0016
Best Alg.	HBOS	*k*-NN	LOF	LoOP	LoOP	*k*-NN	HBOS	HBOS	COF	HBOS

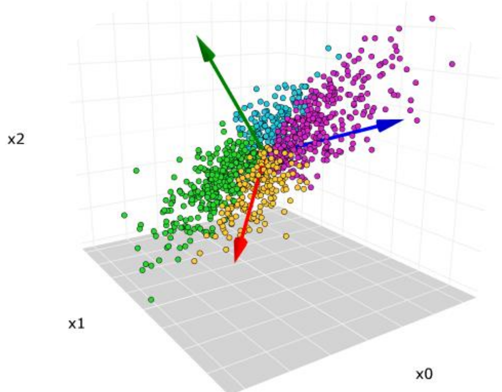
doi:10.1371/journal.pone.0152173.t004

Performance

Suggested Modification

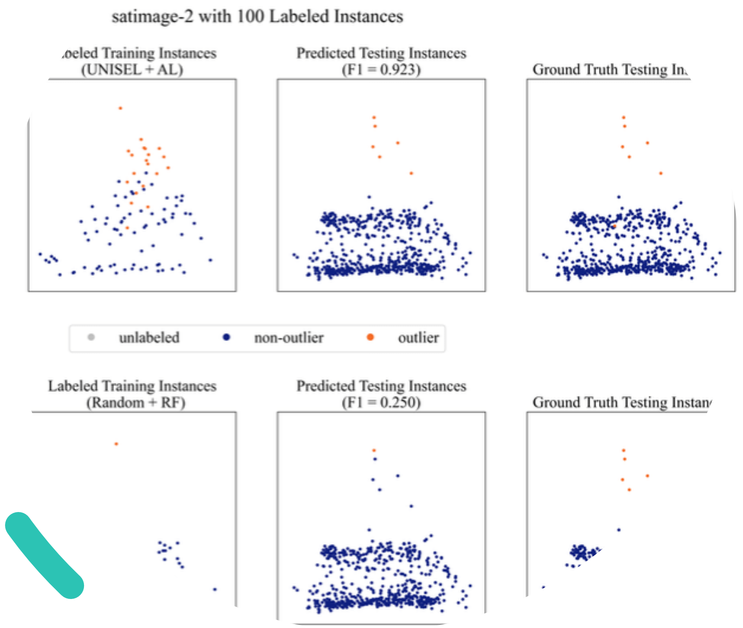
- In order to improve the performance of HBOS while maintaining low complexity, we've tested
  - Running PCA as correlation between features offset result
  - Test softmax on bins
  - Weighting the bins per variance of their features

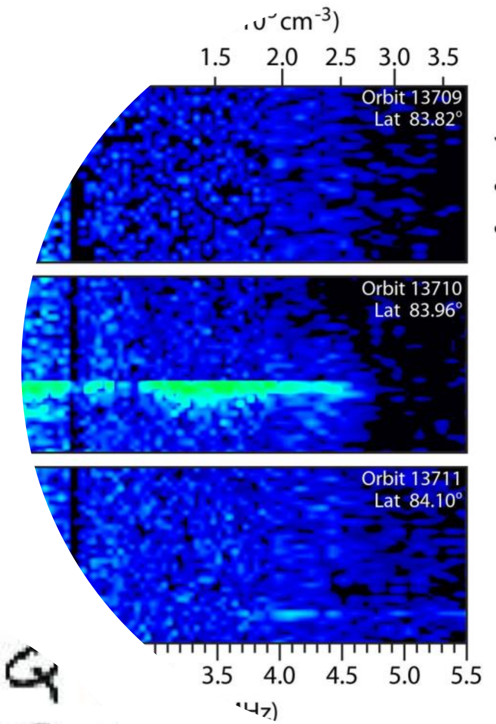




Datasets

- ionosphere - Classification of radar returns from the ionosphere
- letter - For recognizing handwritten forms
- mnist - a large dataset of handwritten digits
- satimage-2 - The original Statlog (Landsat Satellite) dataset





Results

Table 10 Modified HBOS with PCA (M-HBOS) RoC comparison

	Data	#Samples	# Dimensions	Outlier Perc	HBOS	PCA	M-HBOS
0	ionosphere	351	33	35.8974	0.5154	0.8068	0.9601
0	letter	1600	32	6.25	0.5783	0.511	0.804
0	mnist	7603	100	9.2069	0.5775	0.8565	0.7812
0	satimage-2	5803	36	1.2235	0.9864	0.9842	0.9783

Table 11 Modified HBOS with PCA (M-HBOS) Precision comparison

	Data	#Samples	# Dimensions	Outlier Perc	HBOS	PCA	M-HBOS
0	ionosphere	351	33	35.8974	0.3585	0.6226	0.8679
0	letter	1600	32	6.25	0.1	0.1	0.275
0	mnist	7603	100	9.2069	0.1259	0.3741	0.3777
0	satimage-2	5803	36	1.2235	0.7273	0.8485	0.8125

Table 12 Modified HBOS with PCA run time comparison

	Data	#Samples	# Dimensions	Outlier Perc	HBOS	PCA	M-HBOS
0	ionosphere	351	33	35.8974	0.8202	0.0871	0.6102
0	letter	1600	32	6.25	0.0116	0.0114	0.029
0	mnist	7603	100	9.2069	0.0595	0.1433	0.0581
0	satimage-2	5803	36	1.2235	0.02	0.0173	0.0159

Summary

We have presented a suggested improvement to the widely used HBOS algorithm, adding PCA prior the HBOS calculation.

We further implemented the above on 4 commonly used dataset namely ionosphere, letter, mnist and satimage-2 to show significant improvements while maintaining the low complexity.

Thank you

Roy Kinamon & Eliran Elisha