Generalized Anomaly Detection

https://arxiv.org/pdf/2110.15108.pdf

Михеев Борис

ВМК МГУ
Кафедра ММП

22 декабря 2021

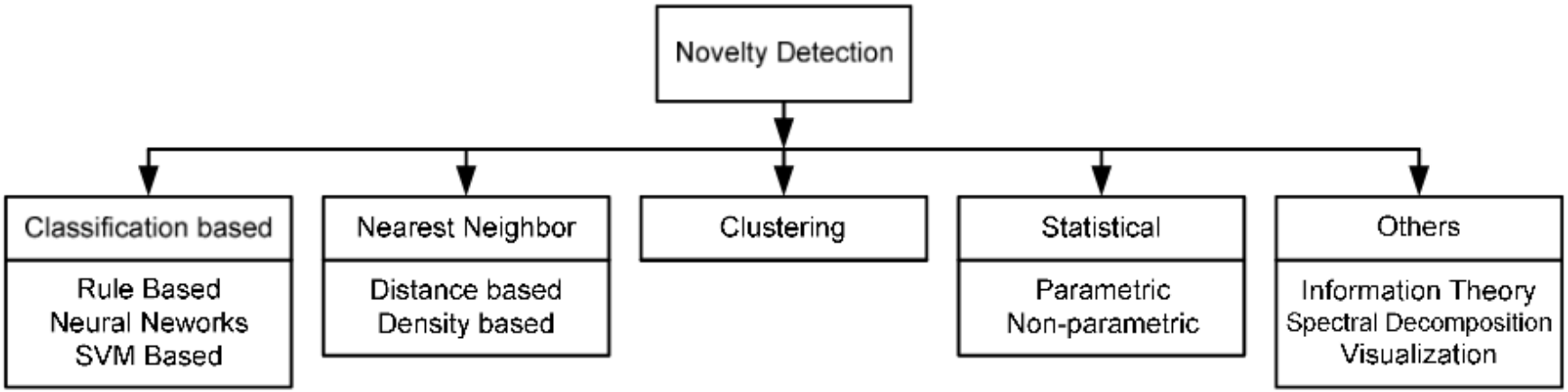
Задача детекции аномалий

Заключается в выявлении новых, некорректных или до сих пор неизвестных модели данных, отличных от типичных точек.
Обычно предполагается, что нормальные данные сосредоточены в компактной области признакового пространства, схожи и близки друг к другу. Объекты вне этой области считаются аномалиями.

О терминологии

- Anomaly - объект из конкретно другого распределения, нет схожести с нормальными объектами
- Outlier - редкий, маловероятный объект, схожий с нормальными
- Novelty - объект из новой области значений вероятности, данные, с которыми модель еще не сталкивалась.

Методы детекции аномалий



Методы, основанные на классификации

В случае, когда нормальный класс состоит из одной категории объектов, классификатор обучается на данных с двумя классами - нормальным и аномальным. Ищется отображение $\mathcal{F}$ из пространства признаков входных объектов $\mathcal{X}$ в признаковое пространство $\mathcal{Y}$ меньшей размерности. Посредством некой функции $\mathcal{D}$ представлениям объектов присваивается степень принадлежности к нормальному/аномальному классу.
Способы обобщения на случай m категорий нормального класса:

- Обучить m классификаторов для каждого класса на своем классе, применять их в совокупности для идентификации нового объекта
- Использовать один классификатор, приняв все нормальные категории за одну

Теория Демпстера-Шаффера

$\mathbb{H} = \{\mathbb{H}_1, ..., \mathbb{H}_n\}$ - множество взаимоисключающих утверждений,
$2^{\mathbb{H}} = \{A | A \subseteq \mathbb{H}\}$
Массовоые функции: $m(\emptyset) = 0$, $m(A) \geq 0$, $\sum\limits_{A|A \subseteq 2^{\mathbb{H}}} m(A) = 1$

Функция доверия: $Bel(A) = \sum\limits_{S|S \subseteq A} m(S) = m(A) + \sum\limits_{S|S \subset A} m(S)$

$Bel(\overline{A}) = \sum\limits_{S|S \cap A = \emptyset} m(S)$

Функция правдоподобия: $Pl(A) = 1 - Bel(\overline{A}) = \sum\limits_{S|S \cap A \neq \emptyset} m(S)$

Правило Демпстера: пусть $m_1(S)$, $m_2(S)$ - два набора масс для всех $S \subseteq \mathbb{H}$
Тогда $m(A) = \frac{1}{\mathcal{N}} \sum\limits_{S_1 \cap S_2 = A} m_1(S_1) m_2(S_2)$, $A \neq \emptyset$

где $\mathcal{N} = \sum\limits_{S \subseteq \mathbb{H}} \sum\limits_{S_1 \cap S_2 = S} m_1(S_1) m_2(S_2) = \sum\limits_{S_1 \cap S_2 \neq \emptyset} m_1(S_1) m_2(S_2) =$
$=1- \sum\limits_{S_1 \cap S_2 = \emptyset} m_1(S_1) m_2(S_2)$

Многоклассовая задача

	$\{1\}$	$\{2\}$	$\cdots$	$\{m\}$	$\neg\{1\}$	$\neg\{2\}$	$\cdots$	$\neg\{m\}$
$(\mathcal{F}_1, D_1)$	$P_1(1\|x)$	0	$\cdots$	0	$1 - P_1(1\|x)$	0	$\cdots$	0
.	.	.	$\cdots$	.	.	.	$\cdots$	.
$(\mathcal{F}_m, D_m)$	0	0	$\cdots$	$P_m(m\|x)$	0	0	$\cdots$	$1 - P_m(m\|x)$

Table 1: Probability estimates for sample $x \in \mathcal{X}$ provided by m classifiers.

$$U = \{1, 2, ..., m, \Lambda\}$$

$$\mathbb{P}(\Lambda|x) = \frac{1}{K} \sum_{\neg\{1\}\cap...\cap\neg\{m\}=\Lambda} \prod_{i=1}^{m} \mathbb{P}_i(\neg\{i\}|x) = \frac{1}{K} \prod_{i=1}^{m} \mathbb{P}_i(\neg\{i\}|x)$$

$$K = 1 - \sum_{u_1\cap...\cap u_m=\emptyset} \prod_{i=1}^{m} \mathbb{P}_i(u_i|x),\ u_i = \{\{i\}, \neg\{i\}\}$$

$$K = \prod_{i=1}^{m} \mathbb{P}_i(\neg\{i\}|x) + \sum_{i=1}^{m} \mathbb{P}_i(\{i\}|x) \prod_{j=1, j\neq i}^{m} \mathbb{P}_j(\neg\{j\}|x)$$

Алгоритмы

Algorithm 1: Using m Single-Class Anomaly Detection for Multi-Class Case.

Training: Let $(\mathcal{F}, D)$ be any one-class anomaly detection algorithm. Given training set $X = \{X_1, X_2, \cdots, X_m\}$ consisting of training examples from m classes $X_i \subset \mathcal{X}_i$, train $(\mathcal{F}, D)$ separately on each class producing m classifiers, $(\mathcal{F}_i, D_i) 1 \leq i \leq m$.

Testing: Given $x \in \mathcal{X}$, classify it with each of the m classifiers. *Declare x anomalous if all classifiers classify it as anomalous.*

Notes: A classifier $(\mathcal{F}, D)$ typically uses a comparison $D(\mathcal{F}(x)) < T$ for classification (T is an arbitrary threshold that can be tuned). For SVDD this is a distance from the center while for DROCC it is distance from a manifold.

Algorithm 2: Training a *single* classifier by combining all m classes.

Training: Given a training set $X = X_1 \cup X_2 \cup \cdots \cup X_m$, train a single classifier $(\mathcal{F}, D)$ on this set (see *Notes* in Algorithm 1).

Testing: An example x is classified as normal or anomalous by this classifier (just like any single-class classifier).

Notes: This algorithm can be applied directly to the *inseparable* model described previously.

Алгоритмы

Lemma 3.1: Algorithm 2 has a higher AUC value than Algorithm 1 when the same single-class anomaly detection algorithm is used in both cases.

Proof: Recall that the AUC value is the integral of the True Positive Rate (TPR) vs False Positive Rate (FPR) curve (each point on the curve corresponds to a different value for the detection threshold T). For Algorithm 1 we can write the TPR and FPR as,

$$\begin{aligned}
\text{TPR}_1 &= 1 - \text{False Negative Rate} \\
&= 1 - \frac{1}{K}\sum_{x_j \in \cup_{i=1}^{m} \mathcal{X}_i} p(x_j) \prod_{i=1}^{m} (1 - P_i(\{i\}|x_j)) \\
\text{FPR}_1 &= 1 - \text{True Negative Rate} \\
&= 1 - \frac{1}{K}\sum_{x_j \in \mathcal{X} \setminus \cup_{i=1}^{m} \mathcal{X}_i} p(x_j) \prod_{i=1}^{m} (1 - P_i(\{i\}|x_j))
\end{aligned}$$

Note that the difference in the two expressions is in the sets that x_j is selected from. The similar expressions for Algorithm 2 are,

$$\begin{aligned}
\text{TPR}_2 &= 1 - \text{False Negative Rate} \\
&= 1 - \frac{1}{K}\sum_{x_j \in \cup_{i=1}^{m} \mathcal{X}_i} p(x_j)(1 - P(\text{Normal}|x_j)) \\
&= 1 - \frac{1}{K}\sum_{x_j \in \cup_{i=1}^{m} \mathcal{X}_i} p(x_j) \sum_{i=1}^{m} \frac{1}{m}(1 - P_i(\{i\}|x_j)) \\
\text{FPR}_2 &= 1 - \text{True Negative Rate} \\
&= 1 - \frac{1}{K}\sum_{x_j \in \mathcal{X} \setminus \cup_{i=1}^{m} \mathcal{X}_i} p(x_j)(1 - P(\text{Normal}|x_j) \\
&= 1 - \frac{1}{K}\sum_{x_j \in \mathcal{X} \setminus \cup_{i=1}^{m} \mathcal{X}_i} p(x_j) \sum_{i=1}^{m} \frac{1}{m}(1 - P_i(\{i\}|x_j))
\end{aligned}$$

The difference in the TPR (and FPR) values of this set of expressions is that for Algorithm 1 we take a product of the form $\prod_i (1 - P_i(\{i\}|x_j))$ whereas for Algorithm 2 this is a mean of the same probabilities. As m increases, we would expect the product term to decrease rapidly. Indeed, the values of TPR and FPR for Algorithm 1 approach 1 as m increases resulting in a purely random classifier. On the other hand, the TPR and FPR for Algorithm 2 remain relatively unchanged since they use an arithmetic mean of the probability rather than a geometric mean as in Algorithm 1. QED

Corollary 3.1.1: As m increases, the AUC value for Algorithm 1 decreases.

Deep AD methods

- Метрические методы
- Основывающиеся на трансформациях
- Генеративные методы

DeepSVDD

Минимизируемая функция потерь:

$$\min_{R,\mathcal{W}} \quad R^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\max\{0, \|\phi(\boldsymbol{x}_i;\mathcal{W}) - \boldsymbol{c}\|^2 - R^2\} + \frac{\lambda}{2}\sum_{\ell=1}^{L}\|\boldsymbol{W}^\ell\|_F^2. \tag{3}$$

DROCC

Algorithm 1 Training neural networks via DROCC

Input: Training data $D = [x_1, x_2, \dots x_n]$.

Parameters: Radius r, $\lambda \geq 0$, $\mu \geq 0$, step-size η, number of gradient steps m, number of initial training steps n_0.

Initial steps: For $B = 1, \dots n_0$

X_B: Batch of training inputs

$$\theta = \theta - \text{Gradient-Step}\Big(\sum_{x \in X_B} \ell(f_\theta(x), 1)\Big)$$

DROCC steps: For $B = n_0, \dots n_0 + N$

X_B: Batch of training inputs

$\forall x \in X_B : h \sim \mathcal{N}(0, I_d)$

Adversarial search: For $i = 1, \dots m$

1. $\ell(h) = \ell(f_\theta(x + h), -1)$
2. $h = h + \eta \frac{\nabla_h \ell(h)}{\|\nabla_h \ell(h)\|}$
3. $h = \frac{\alpha}{\|h\|} \cdot h$ where $\alpha = r \cdot \mathbb{1}[\|h\| \leq r] + \|h\| \cdot \mathbb{1}[r \leq \|h\| \leq \gamma \cdot r] + \gamma \cdot r \cdot \mathbb{1}[\|h\| \geq \gamma \cdot r]$

$$\ell^{itr} = \lambda \|\theta\|^2 + \sum_{x \in X_B} \ell(f_\theta(x), 1) + \mu \ell(f_\theta(x + h), -1)$$

$$\theta = \theta - \text{Gradient-Step}(\ell^{itr})$$

$$\ell^{\text{dr}}(\theta) = \lambda \|\theta\|^2 + \sum_{i=1}^{n} [\ell(f_\theta(x_i), 1) + \mu \max_{\substack{\tilde{x}_i \in \\ N_i(r)}} \ell(f_\theta(\tilde{x}_i), -1)],$$

$$N_i(r) \stackrel{\text{def}}{=} \Big\{ \|\tilde{x}_i - x_i\|_2 \leq \gamma \cdot r; \;\; r \leq \|\tilde{x}_i - x_j\|, \quad \forall j = 1, 2, \dots n \Big\}, \qquad (1)$$

DeepMAD

Algorithm 3: DeepMAD

Training: Randomly initialize m autoencoders A_i;
For every A_i, train A_i on provided examples X_i;
Using the encoder part E_i of the autoencoder, identify a point c_i, the "center" for this class;
For every encoder E_i, create *labeled* training data $\{(x, l)|$ if $x \in X_i$ then $l = +1$ else if $x \in \bigcup_{j=1, j\neq i}^{m} X_j$ then $l = -1\}$.
Then train E_i on this data using loss function $\mathcal{L}$
Result: m trained encoders E_i

Testing: Given x to classify,
Compute $d(x) = \min_{i=1}^{m} ||c_i - E_i(x; \theta_i)||_2$;
If $d(x) < \gamma$ then x is *normal* else x is anomalous

$$\mathcal{L}(\theta_i) = \frac{1}{N} \sum_{j=1}^{N_i} ||E_i(x_j; \theta_i) - c_i||^2 +$$
$$\frac{\eta}{N} \sum_{x_k \in (\bigcup_j X_j) \setminus X_i} (\max(0, \delta - ||E_i(x_k; \theta_i) - c_i||_2)^2)$$
$$+ \frac{\lambda}{2} \sum_{l=1}^{L} ||W^l||_F^2$$

Эксперименты

	2-in, 8-out	5-in, 5-out	9-in, 1-out
DROCC	$0.4728 \longleftrightarrow 0.7252$	$0.4316 \longleftrightarrow 0.7219$	$0.4107 \longleftrightarrow 0.7146$
	$\pm 0.0119 \quad \pm 0.0081$	$\pm 0.0257 \quad \pm 0.0039$	$\pm 0.0454 \quad \pm 0.0079$
Outlier	$(0, 8) \quad 0.8359 \pm 0.0117$		
DROCC(m)	$0.4216 \longleftrightarrow 0.6912$	$0.3806 \longleftrightarrow 0.7023$	$0.3439 \longleftrightarrow 0.6896$
	$\pm 0.0424 \quad \pm 0.0188$	$\pm 0.0047 \quad \pm 0.0648$	$\pm 0.1034 \quad \pm 0.0453$
Outlier	$(0, 8) \quad 0.8255 \pm 0.0137$		
DeepSVDD	$0.4088 \longleftrightarrow 0.7623$	$0.3382 \longleftrightarrow 0.7105$	$0.3058 \longleftrightarrow 0.6844$
	$\pm 0.0068 \quad \pm 0.0193$	$\pm 0.0076 \quad \pm 0.0077$	$\pm 0.0169 \quad \pm 0.0144$
DeepSVDD(m)	$0.4147 \longleftrightarrow 0.7516$	$0.3482 \longleftrightarrow 0.6909$	$0.3580 \longleftrightarrow 0.5864$
	$\pm 0.0129 \quad \pm 0.0093$	$\pm 0.0123 \quad \pm 0.0133$	$\pm 0.0166 \quad \pm 0.0167$
DeepMAD	$0.5396 \longleftrightarrow 0.7647$	$0.4929 \longleftrightarrow 0.7738$	$0.5437 \longleftrightarrow 0.7230$
	$\pm 0.0031 \quad \pm 0.0014$	$\pm 0.0046 \quad \pm 0.0022$	$\pm 0.0028 \quad \pm 0.0084$

Table 2: AUC range for CIFAR-10

	fMNIST			RECYCLE		CIFAR-100
	2-in, 8-out	5-in, 5-out	9-in, 1-out	4-in, 1-out	5-in, 1-out	2-in, 18-out
DROCC	$0.6873 \leftrightarrow 0.9774$	$0.5738 \leftrightarrow 0.9260$	$0.5408 \leftrightarrow 0.8247$	$0.4447 \leftrightarrow 0.7997$	90.56	$0.3548 \leftrightarrow 0.7329$
	$\pm 0.0937 \pm 0.0049$	$\pm 0.0397 \pm 0.0307$	$\pm 0.0961 \pm 0.0507$	$\pm 0.0176 \pm 0.0719$		$\pm 0.0006 \pm 0.0971$
Mean	0.8161	0.7448	0.6992	0.6128		0.5638
Deep SVDD	$0.6622 \leftrightarrow 0.9871$	$0.5438 \leftrightarrow 0.9279$	$0.4551 \leftrightarrow 0.8825$	$0.3703 \leftrightarrow 0.8728$	90.12	$0.4196 \leftrightarrow 0.7185$
	$\pm 0.0502 \pm 0.0033$	$\pm 0.0274 \pm 0.0325$	$\pm 0.0285 \pm 0.0137$	$\pm 0.0207 \pm 0.0079$		$\pm 0.0077 \pm 0.0180$
Mean	**0.8538**	0.7269	0.6523	0.5791		0.5559
Deep MAD	$0.6434 \leftrightarrow 0.9714$	$0.5732 \leftrightarrow 0.8832$	$0.4860 \leftrightarrow 0.9395$	$0.5906 \leftrightarrow 0.8283$	**98.38**	$0.5384 \leftrightarrow 0.8213$
	$\pm 0.0640 \pm 0.0011$	$\pm 0.0485 \pm 0.0137$	$\pm 0.0267 \pm 0.3466$	$\pm 0.0035 \pm 0.0073$		$\pm 0.0018 \pm 0.0012$
Mean	0.8329	**0.7739**	**0.7613**	**0.6966**		**0.6580**

Table 3: AUC range for fMNIST, RECYCLE, and CIFAR-100

Эксперименты

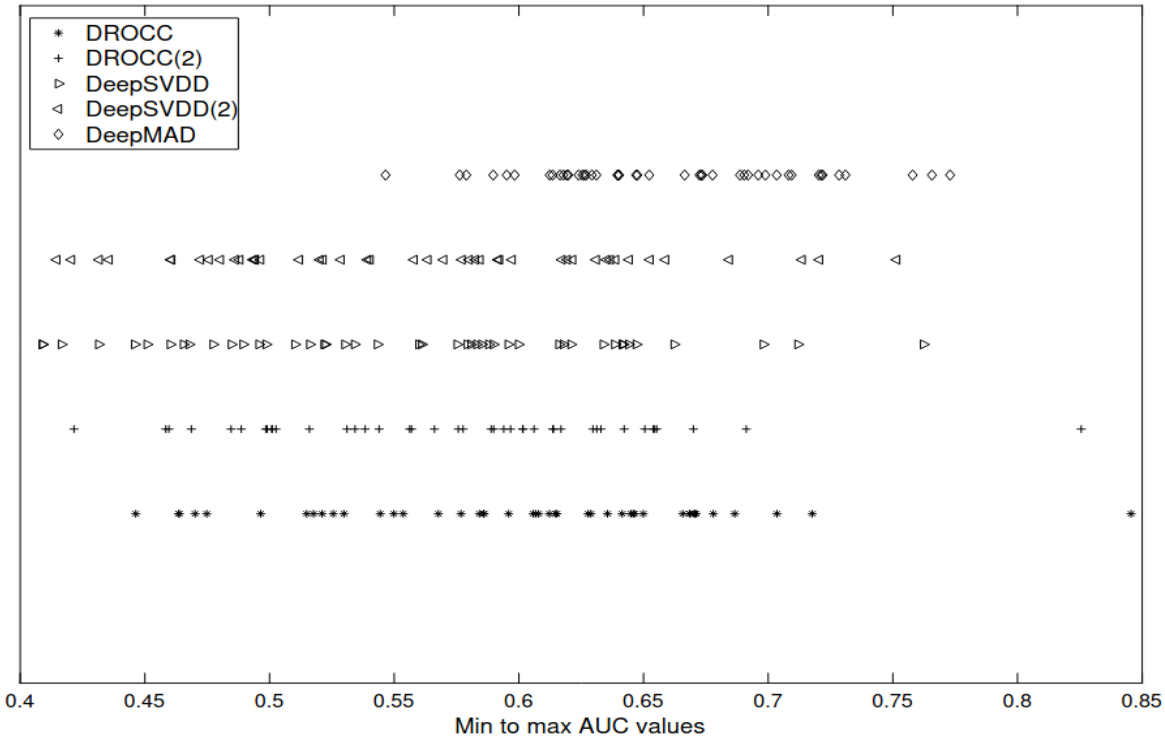


Figure 3: 2 in and 8 out case (CIFAR-10).

Эксперименты

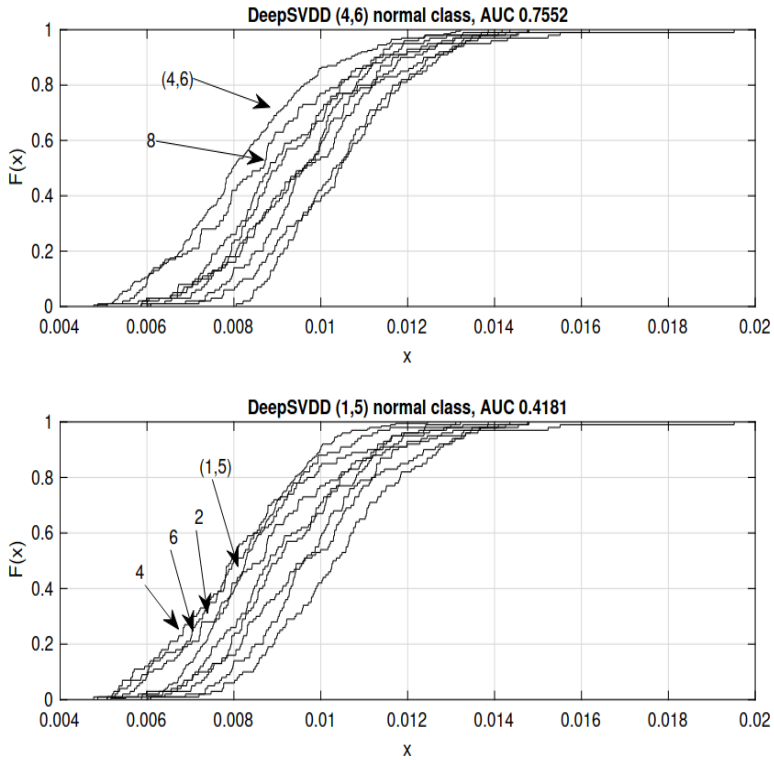


Figure 5: CDF of distances DeepSVDD.

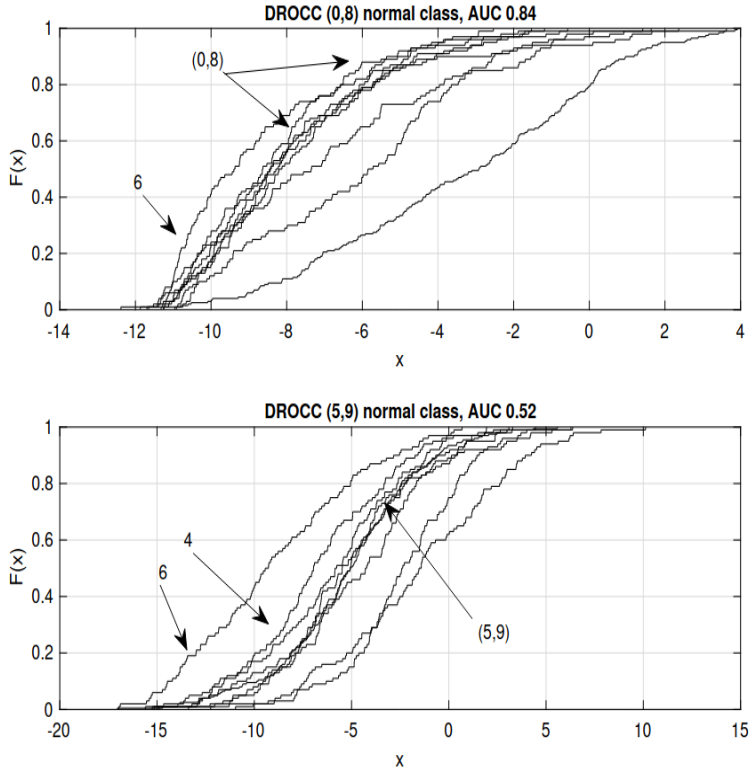


Figure 6: CDF of logits output by DROCC.

Эксперименты

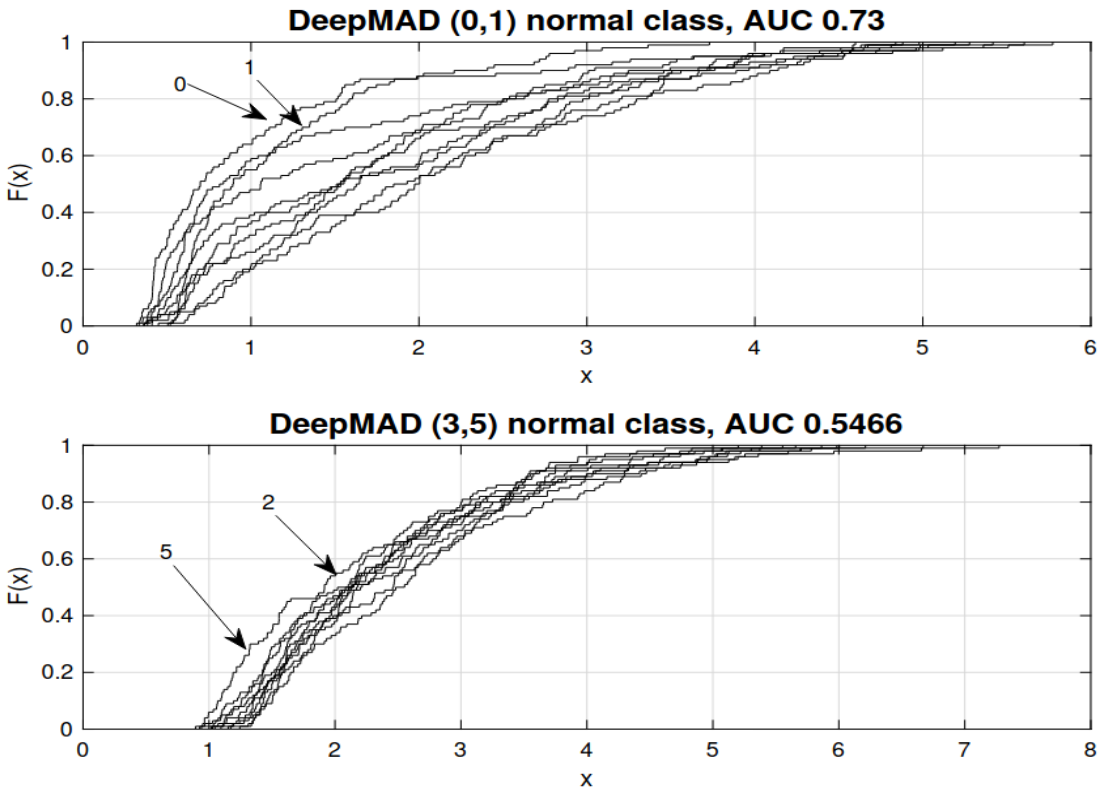


Figure 7: CDF of distances from ‘center’ DeepMAD.

Эксперименты

DROCC									
0	1	2	3	4	5	6	7	8	9
KS((0,8),-), AUC = 0.84									
0	63	57	61	61	61	64	63	0	63
KS((5,9),-), AUC = 0.5									
45	12	36	50	51	14	52	31	38	14
DeepSVDD									
KS((4,6),-), AUC = 0.75									
37	45	9	30	0	24	0	29	36	49
KS((1,5),-), AUC = 0.41									
28	0	18	13	18	0	34	15	31	14
DeepMAD									
In classes (0,1), AUC=0.7174									
KS(0,-)									
0	100	44	127	124	128	126	128	60	125
KS(1,-)									
100	0	101	123	118	123	123	121	107	124
In classes (3,5), AUC=0.5277									
KS(3,-)									
6	1	4	0	16	96	2	73	81	99
KS(5,-)									
96	94	77	96	87	0	97	29	105	106

Table 4: h values for the KS test.