Anomaly Detection in Neural Networks via One-Class Support Vector Methods

Poorna Sandamini Senaratne
Advisor: Dr. Edgard Maboudou

Department of Statistics and Data Science
University of Central Florida

August 7, 2025

Contents

One-Class Classification (OCC)

- Binary classification involves predicting class labels for observations from two classes using a training set with known labels.
- One-class classification, a special case of binary classification, deals with data from a single known class (target) without information about other possible classes.
- It was introduced by Moya et al.,1993 [4] using neural networks to create boundaries around the target class data.

Applications of OCC

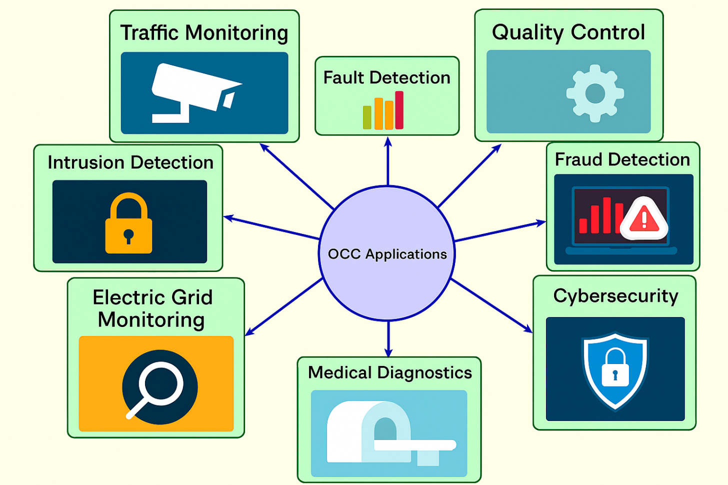

Figure 1: Some applications of OCC

Background on One-Class Classification

- Support Vector Data Description (SVDD), introduced by Tax and Duin in 2004 [5], is a kernel-based boundary method used for one-class classification and novelty detection.
- Least Squares Support Vector Data Description (LS-SVDD), proposed by Guo et al. in 2017 [2], modifies SVDD using a squared error loss with equality constraints.
- Maboudou-Tchao (2021) [3] showed that LS-SVDD has a closed-form solution, making it computationally attractive.

Visualization of LS-SVDD

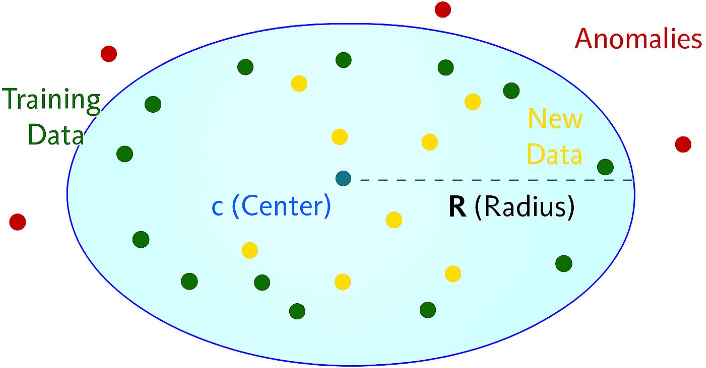

Figure 2: LS-SVDD Visualization

What is a Neural Network?

- A neural network is a computational model inspired by the human brain that consists of interconnected neurons (nodes).
- It consists of an input layer, hidden layers, and an output layer.
- Each neuron applies a mathematical transformation to input data and passes the output to the next layer.

Neural Network Architecture

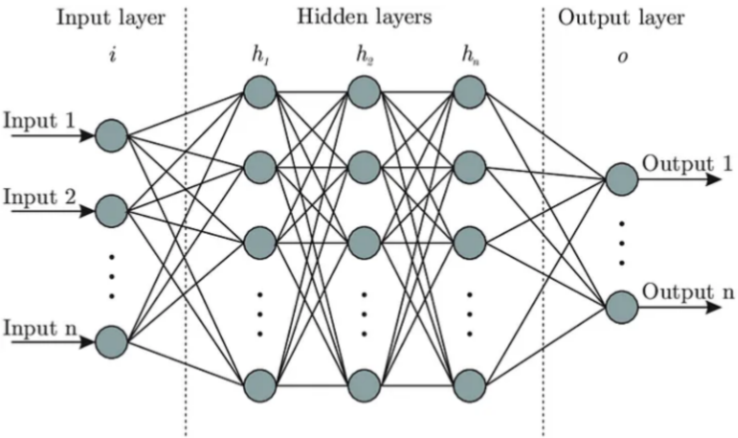

Figure 3: Architecture of a multilayer neural network with three hidden layers

Embedding Layer

- In artificial neural networks, an embedding layer transforms input data into a dense, low-dimensional representation called an embedding, capturing the most relevant features for the task at hand.
- Embeddings are commonly used to compress the information while preserving important relationships in the data, making them suitable for tasks such as classification and anomaly detection.
- In one-class classification, embeddings generated by neural networks can be used to detect outliers or anomalies in the neural network paramters or in the data distribution.

Embedding Layer

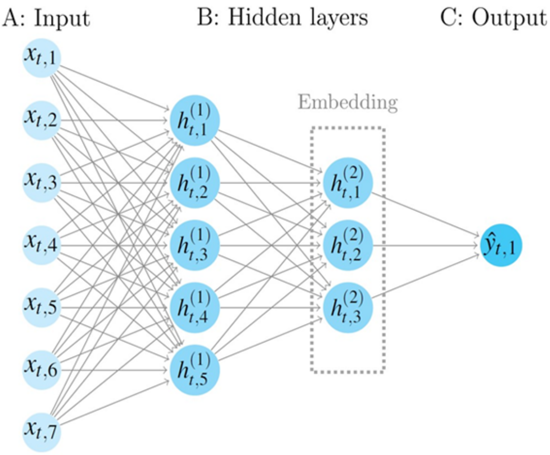

Figure 4: The FNN with two hidden layers in toy example, Malinovskaya et al., 2024 [1]

Proposed Framework: Overview

Embedded Least Squares Support Vector Data Description (ELS-SVDD)

- We combine neural network embeddings with LS-SVDD for effective one-class classification.
- Embeddings capture complex data patterns; LS-SVDD provides a boundary in transformed space.
- The method is simple, scalable, and offers a closed-form computation.

Proposed Framework: Training Phase (NN + ELS-SVDD)

Neural Network Training

- Let $\mathcal{D}_N = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subseteq \mathcal{X}$ denote a training dataset, where $\mathcal{X} \subseteq \mathbb{R}^p$ is the input feature space and N is the total number of samples.
- We associate this input space with an output space $\mathcal{F} \subseteq \mathbb{R}^d$.
- Consider a neural network $\phi(\cdot; \mathcal{W}, \mathcal{B})$ with $L \in \mathbb{N}$ layers that defines a mapping from $\mathcal{X}$ to $\mathcal{F}$.
- The network comprises L hidden layers and one output layer, where $\mathcal{W} = \{\mathbf{W}^{[1]}, \mathbf{W}^{[2]}, \ldots, \mathbf{W}^{[L]}\}$ represents the set of weight matrices, and $\mathcal{B} = \{\mathbf{b}^{[1]}, \mathbf{b}^{[2]}, \ldots, \mathbf{b}^{[L]}\}$ denotes the corresponding bias vectors.

Training Phase cont.

- We get to the final hidden layer $\ell = L - 1$. For each sample $i \in \{1, 2, \ldots, N\}$, we compute

$$\mathbf{z}_i^{[\ell]} = \mathbf{W}^{[\ell]}\mathbf{a}_i^{[\ell-1]} + \mathbf{b}^{[\ell]} \tag{1}$$

- Let $\mathbf{u}_i = \mathbf{z}_i^{[L-1]} \in \mathbb{R}^m, \quad i = 1, \ldots, N$ denote the embedding vector extracted from the $\ell = L - 1$ hidden layer for sample $\mathbf{x}_i$, where m ($m < p$) is the number of neurons in the selected embedding layer (i.e., the dimensionality of each latent vector).
- Collecting all N such embeddings into a single matrix, we define the embedding matrix $\mathbf{U} \in \mathbb{R}^{N\times m}$ as, $\mathbf{U} = \begin{bmatrix}\mathbf{u}_1^\top & \mathbf{u}_2^\top & \cdots & \mathbf{u}_N^\top\end{bmatrix}^\top$

Training Phase cont.

ELS-SVDD Training

- Let the embeddings from the neural network, be $\{\mathbf{u}_i\}_{i=1}^N$, $\mathbf{u}_i \in \mathbb{R}^m$. Then the optimization problem can be denoted as:

$$\min_{R,a,\xi} \; R^2 + \frac{C}{2}\sum_{i=1}^{N} \xi_i^2 \tag{2}$$

subject to:

$$\|\varphi(\mathbf{u}_i) - \mathbf{a}\|^2 = R^2 + \xi_i, \quad i = 1, 2, \ldots, N$$

where:

  - R: radius of the hypersphere
  - $\mathbf{a}$: center of the hypersphere
  - C: $C > 0$ is introduced to control the influence of the error variables
  - ξ: error variables realized by a training vector $\mathbf{u}_i$ with respect to the hypersphere
  - $\varphi(\mathbf{u}_i)$: the mapping of $\mathbf{u}_i$ to a higher-dimensional feature space

Training Phase cont.

ELS-SVDD Solution

Given embeddings $\mathbf{U}$ and a kernel function $k(\cdot,\cdot)$, satisfying the Mercer's theorem, the optimal support vector coefficients $\boldsymbol{\alpha}_u$ that minimizes the ELS-SVDD objective is given in closed form by:

$$\boldsymbol{\alpha}_u = \frac{1}{2}\mathbf{H}_u^{-1}\left(\mathbf{k}_u + \frac{2 - \mathbf{e}^T\mathbf{H}_u^{-1}\mathbf{k}_u}{\mathbf{e}^T\mathbf{H}_u^{-1}\mathbf{e}}\mathbf{e}\right) \tag{3}$$

where

- $\mathbf{K}_u \in \mathbb{R}^{N\times N}$: Gram matrix with entries $[\mathbf{K}_u]_{ij} = k(\mathbf{u}_i, \mathbf{u}_j)$
- $\mathbf{H}_u = \mathbf{K}_u + \frac{1}{2C}\mathbf{I}_N$: regularized kernel matrix
- $\mathbf{k}_u = [k(\mathbf{u}_1, \mathbf{u}_1), \ldots, k(\mathbf{u}_N, \mathbf{u}_N)]^T$
- $\mathbf{e}$: N-dimensional vector of ones

Training Phase cont.

Radius of ELS-SVDD

- The squared radius of the hypersphere enclosing the embedded data points is given by:

$$R_u^2 = \frac{1}{N}\sum_{s=1}^{N}\left(k(\mathbf{u},\mathbf{u}) - 2\sum_{i=1}^{N}\alpha_{u\,i}k(\mathbf{u},\mathbf{u}_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{u\,i}\alpha_{u\,j}k(\mathbf{u}_i,\mathbf{u}_j)\right) \tag{4}$$

Proposed Framework: Inference Phase

- Given a new test input $\mathbf{x}^* \in \mathbb{R}^p$, we compute its latent representation of the final hidden layer ($\ell = L - 1$) using the trained neural network. Let's denote this latent representation as $\mathbf{v}^* \in \mathbb{R}^m$:

$$\mathbf{v}^* = \mathbf{W}^{[\ell]}\mathbf{a}^{[\ell-1]}(\mathbf{x}^*) + \mathbf{b}^{[\ell]} \tag{5}$$

- We then compute its squared distance from the hypersphere center in kernel space:

$$d_{\mathbf{v}^*} = k(\mathbf{v}^*, \mathbf{v}^*) - 2\sum_{j=1}^{N} \alpha_j k(\mathbf{v}^*, \mathbf{u}_j) + \sum_{j=1}^{N}\sum_{k=1}^{N} \alpha_j \alpha_k k(\mathbf{u}_j, \mathbf{u}_k) \tag{6}$$

Inference Phase cont.

Decision Rule of ELS-SVDD

- The test point is then classified according to the rule:

$$\text{Class}(\mathbf{x}^*) = \begin{cases} \text{Target}, & \text{if } d_{\mathbf{v}^*} \leq R_u^2 \\ \text{Outlier}, & \text{otherwise} \end{cases} \tag{7}$$

Dataset Overview: Internet Firewall Data

- Source: UCI Machine Learning Repository
- Collection: Internet traffic records captured from a university's firewall
- Type: Multivariate Classification dataset
- Instances: 65,532
- Features: 12
- Target Variable: `Action`
  - Classes: `allow`, `action`, `drop`, `reset-both`

Steps

- We select 500 observations for training and 100 for testing.
- This corresponds to p = 12 and N = 500.
- We used the training set to train the ANN, i.e obtain $\mathcal{W}$ and $\mathcal{B}$.
- Next, obtain the embedding vectors, $\mathbf{u}_i \in \mathbb{R}^4$. Here, $m = 4$.
- Apply ELS-SVDD to the embeddings $\mathbf{u}_i$.

Example output

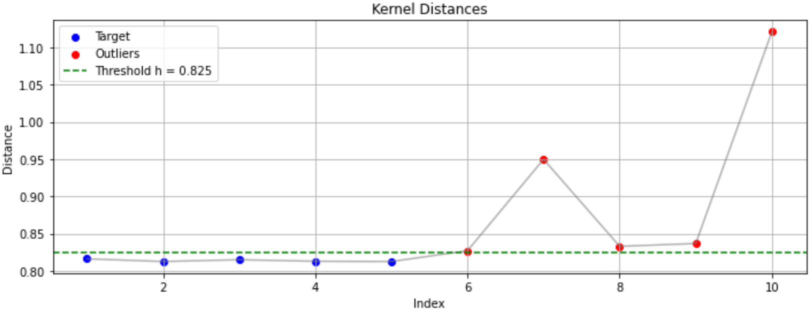

Figure 5: Consecutive outliers

Output Interpretation

- The output exhibits a sequence of consecutive outliers, a cluster of kernel distances that consecutively exceed the threshold.
- In the context of neural networks, consecutive outliers in the embedding space might be indicators of a change in the internal feature representations, implying that the network parameters or input characteristics have shifted significantly.
- This may warrant model retraining or adaptation to ensure continued predictive performance.

Conclusion

- We proposed a framework combining neural network embeddings with Least Squares Support Vector Data Description (ELS-SVDD) for anomaly detection.
- Application to the real-world Internet Firewall dataset further supports the method's effectiveness in identifying distributional changes and potential anomalies.
- This framework provides a principled, data-driven solution for monitoring learned representations in deep learning models.
- Future work: Extend this work to 'Concept drift for streaming data'.

References I

[1] Pavlo Mozharovskyi Anna Malinovskaya and Philipp Otto. "Statistical Process Monitoring of Artificial Neural Networks". In: *Technometrics* 66.1 (2024), pp. 104–117. URL: https://doi.org/10.1080/00401706.2023.2239886.

[2] Yu Guo, Huaitie Xiao, and Qiang Fu. "Least square support vector data description for HRRP-based radar target recognition". In: *Applied Intelligence* 46 (2017), pp. 365–372.

[3] Edgard M Maboudou-Tchao. "Monitoring the mean with least-squares support vector data description". In: *Gestão & Produção* 28 (2021), e019.

[4] Mary M. Moya, Mark W. Koch, and Larry D. Hostetler. "One-class classifier networks for target recognition applications". In: (1993), p. 24043d M. URL: https://api.semanticscholar.org/CorpusID:108681837.

References II

[5] David MJ Tax and Robert PW Duin. "Support vector data description". In: *Machine learning* 54 (2004), pp. 45–66.

Thank you!