FPGA IMPLEMENTATIONS OF KERNEL-BASED ANOMALY DETECTION

Stephen Tridgell
A. Prof. Philip Leong
School of Electrical Engineering
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Why Anomaly Detection?

- **Industrial Monitoring**
  - High value manufacturing equipment wears down over time and anomaly detectors can be used to detect faults in the equipment before it breaks
  - Can also be applied to highly complex systems such as commercial aircraft.
- **Electronic or Network Security**
  - Anomaly detection algorithms are currently used to detect fraudulent credit card transactions
  - Can be applied to network security, detecting events that suggest inappropriate or unauthorized use of a system.
- **Health Care**
  - Can be used to detect unusual sections of scans such as MRIs or PET scanners
  - Monitors patient symptoms and prescribed mediation bringing the attention of a doctor to unusual cases so that timely medical treatment is received.

Theory

- **The Concept of Risk**

$$R(\alpha) = \int \frac{1}{2}|f_\alpha(x) - y| dP(x,y)$$

  - This equation calculates the risk that an example is misclassified. Unfortunately, the distribution that is integrated over is unknown.
  - It can be approximated by using previous examples known as the empirical risk.

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{2}|f_\alpha(x_i) - y_i|$$

  - The relationship between them is:

$$R(\alpha) \leq R_{emp}(\alpha) + \phi(\frac{h}{l})$$

$$\phi\left(\frac{h}{l}\right) = \sqrt{\frac{h(log\frac{2l}{h}+1) - log(\frac{\eta}{4})}{l}}$$

  - As it is not possible to minimize the actual risk, the empirical risk could be minimized instead leading to a classifier with good accuracy on the training data but poor accuracy on unknown data known as *overfitting*
  - The *regularization* term, $\phi\left(\frac{h}{l}\right)$, expresses a preference for a simpler solution, resulting in a better detector

- **The Representer Theorem**
  - Can be summarized as allowing the function f_α to be expressed as a linear combination of examples seen so far
  - In Support Vector Machines (SVMs) these are known as support vectors
  - The representer theorem expresses the boundary or hyperplane as a function of the support vectors such as:

$$f_\alpha(x) = \sum_i \alpha_i \langle x_i, x\rangle + \rho$$

- **The Kernel Trick**
  - Used to provide non-linear separating boundaries efficiently
  - Allows the dot product of the vectors to be replaced with another measure of similarity such as:

$$\langle x_i, x\rangle \rightarrow k(x_i, x) = e^{-\frac{||x_i - x||^2}{\sigma}}$$

Kernel-Based Anomaly Detectors

- **Online vs Batch solvers**
  - Batch algorithms take all the examples at once and try to find the solution
  - Online algorithms update their solution for each new example
  - Batch algorithms generally require large amounts of memory to store all the points simultaneously and hence are not ideal for low latency FPGA applications.

- **One Class SVMS**
  - Scholkopfs One Class SVM is a widely used anomaly detector
  - It trains on examples of only one class (the normal examples)
  - It uses a hyperplane to separate normal and anomalous data
  - This expression is solved to find the optimal separating hyperplane.

$$\min_{\omega,\xi,\rho} \quad \frac{1}{2}||\omega||^2 - \rho + \frac{1}{\nu l}\sum_i \xi_i$$

  - an Online method to solve this optimization problem is the C&P incremental algorithm
  - In this research, it was determined that the C&P algorithm is not ideal for FPGAs due to dependencies and changing sizes of matrices and vectors

- **NORMA**
  - NORMA reduces the instantaneous risk (the risk at the currently added point) and takes a step in the direction that reduces it
  - It is computationally cheap in comparison to the C&P algorithm
  - Each step, the new example is added, the oldest removed and the weights of the rest are reduced

- **OLKn Algorithm**
  - The OLKn algorithm is a recently proposed improvement to NORMA
  - It is formulated as:

$$\min_{\omega_{t+1},\xi_{t+1},\rho_{t+1}} \quad \frac{1}{2}\|\omega_{t+1} - \omega_t\|^2 + \frac{r}{2}\|\omega_{t+1}\|^2 + C\xi_{t+1} - \nu\rho_{t+1}$$

$$s.t. \quad \omega_{t+1}\cdot\Phi(x_{t+1}) - 1 > \rho_{t+1} - \xi_{t+1}$$

$$\xi_{t+1} \geq 0, \rho_{t+1} \geq 0$$

  - it was shown in this research that this algorithm could be theoretically simplified, reducing the complexity of the update
  - The modified algorithm is presented as:

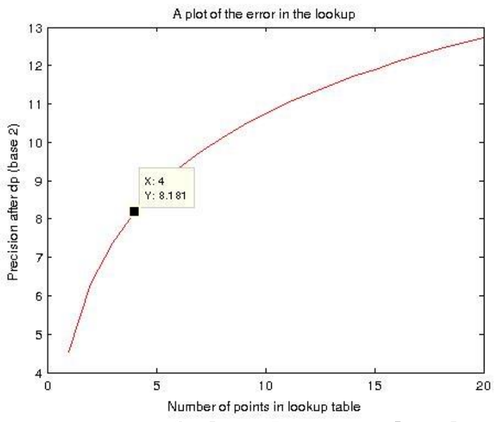
$$\min_{\omega_{t+1},\xi_{t+1}} \quad \frac{1}{2}\|\omega_{t+1} - \omega_t\|^2 + \frac{r}{2}\|\omega_{t+1}\|^2 + C\xi_{t+1}$$

$$s.t. \quad \omega_{t+1}\cdot\Phi(x_{t+1}) - 1 > -\xi_{t+1} \ , \ \xi_{t+1} \geq 0$$

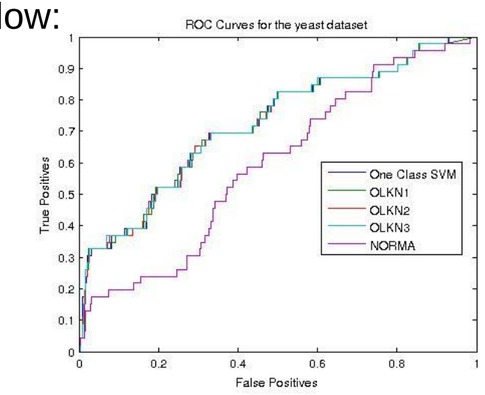
  - Another contribution are the proposed methods for a budgeted OLKn algorithm which are:
    1. Sliding Window
    2. Conditional Sliding Window
    3. Conditional Sliding Window with removal of the minimal weighted vector
  - The first method is similar to NORMA where for each new example, the oldest one is removed
  - The conditional sliding window is a modification where a point is only added if it is considered important enough
  - The last method uses this in combination with removal of the minimum weighted example instead of the oldest

Implementations and Techniques

- **Evaluation of the Gaussian Kernel**
  - The calculation of the Gaussian Kernel is a critical section for any kernel-based machine learning algorithm
  - Computing the exponential is expensive on a FPGA but by:
    - moving to a power of 2 instead of e
    - using a linear interpolation lookup table with fixed point arithmetic
  - significant increases in performance were demonstrated
  - The plot shows the accuracy of a lookup table for different numbers of points

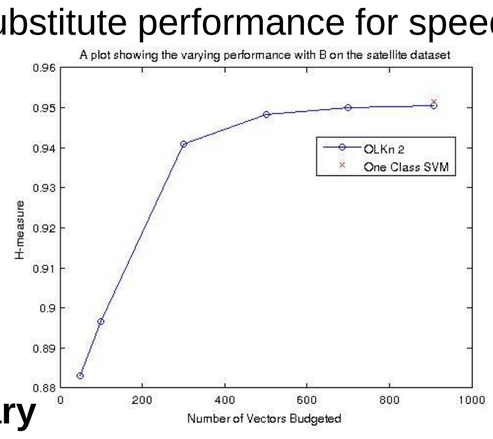


- **Performance of the Anomaly detectors**
  - The performance of various anomaly detectors was evaluated to provide a comprehensive overview of OLKn
  - The results for one of these datasets which involves detecting the type of yeast organism is shown in the figure below:



  - OLKn is slightly inferior to the One Class SVM, but the C&P algorithm has a cost of $O(n^2)$ where as the OLKn algorithm is $O(n)$, a significant reduction in computation and space
  - The lack of dependencies in the OLKn algorithm also makes it ideal for an FPGA application

  - An advantage of budgeted algorithms such as OLKn and NORMA over the One Class SVM is the ability to substitute performance for speed.



Summary

- Investigated anomaly detection algorithms for low latency applications using FPGAs
- The OLKn algorithm was improved and subsequently compared in detail to other anomaly detection algorithms
- Techniques to be used in FPGA applications were proposed such as the computation of the Gaussian Kernel using linear interpolation and a lookup table. These were implemented on a FPGA showing a dramatic improvement in performance.