

# **Project Title: Anomaly Detection with Nonparametric Sequential Probability Ratio Tests**

**CALCE team:** Vasilis A. Sotiris, Shunfeng Cheng and Michael Pecht

## **Objective:**

Develop a general approach for anomaly detection in complex multivariate systems

## **Introduction:**

The motivation for pursuing a nonparametric approach to detection comes from insufficient detection accuracy of traditional linear pattern recognition algorithms. These traditional approaches assume that the underlying distribution of the data is Gaussian or at least mostly Gaussian. The similarity measures that these approaches compute are generally not useful for data that are highly non – Gaussian and covariate dependant. A general detection approach that does not suffer from parametric constraints, whether Gaussian or other imposed distributions (Lognormal, Gamma, etc.) is anticipated to overcome these problems.

Additionally, in order to efficiently analyze high dimensional data a feature extraction approach based on centroid clustering is used to pre – process the data. In the clustering approach data similarity is computed based on Euclidian (Mahalanobis, distances of observations to representative centroids of the training populations. A distribution of the training distances is achieved through Monte Carlo sampling of the training data.

With the use of a nonparametric sequential probability test (NPSVRT), new test distances from new observations are compared to the derived training distribution to infer their classification and in turn the system health.

## **Approach:**

This project studies the application of an unsupervised learning algorithm called the multivariate centroid clustering (MCC) algorithm. The algorithm falls under the category of data clustering algorithms, a common technique for statistical data analysis. Clustering is the classification of similar objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Clustering as a statistical analysis tool is very useful in situations where the data distribution is un-known or non – Gaussian.

The trait extracted with MCC is the Euclidian distance of data to the computed centroid of the data population. A monte calro simulation extracts a univariate density of these distances, which from the training population produce a distribution of healthy distances. A kernel density estimate will provide a continuous form for the density estimate and used as an input to NPSVRT. In the conventional fixed-sample-size case, for a general class of statistical hypotheses genuinely distribution-free (nonparametric) tests exist (whose type I error probability does not depend on the underlying distribution) and within this broad class there exists suitable tests which are (at least, asymptotically or locally) optimal (i.e., most powerful) against specific subclasses of distributions.

## **Deliverables:**

- Provide a general approach for anomaly detection which uses a nonparametric SPRT
- Provide an algorithm to process high dimensional and non – Gaussian data in combination with NPSVRT
- Data analysis using MCC and NPSVRT on experimental data

**Project Status:**

Theory for the multivariate clustering has been finished and posted as a CALCE project: Multivariate Centroid Clustering (MCC). The team is currently researching the nonparametric SPRT theory.

**Estimated Schedule**

The development of the nonparametric SPRT matlab prototype is planned for April 2008. Testing and validation of code is planned for May 2008 and a complete report in June 2008.