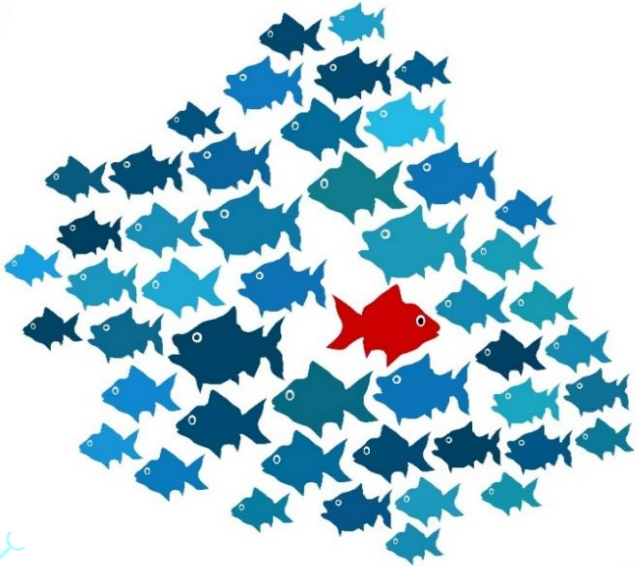
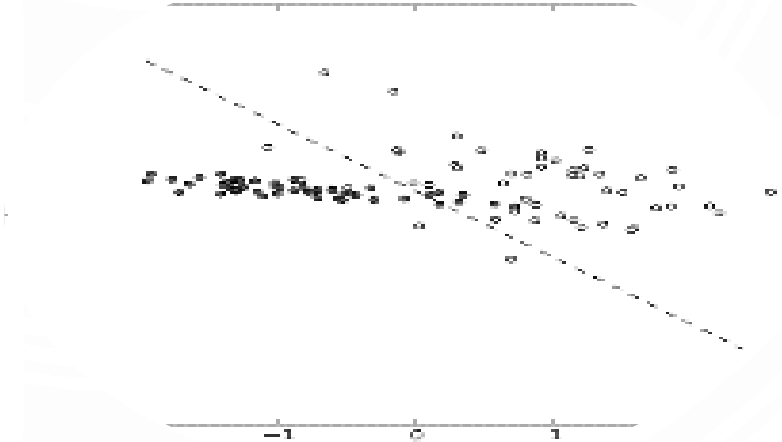
Outlier detection

蔡瑞煌

Data characteristics

- Unlabeled





Data characteristics

14.51, 17.34, 5.33, 6.01, 0.95, 5.63, 0.98, 19.59, 6.52, 1.63, 0.66, 9.53, -0.15, -0.07, 10.16, 12.05, 6.72, 18.47, 17.17, 16.06, 5.24, 6.62, 7.24, 3.57, 17.36, 19.32, 3.79, 1.76, 8.83, 17.06, 11.18, 0.43, 4.76, 13.46, 8.69, 17.47, 20.3, 11.33, 10.33, 5.63, 17.65, -0.11, 3.07, 8.37, 15.08, 17.9, 6.53, 4.77, 10.55, 1.52

Time series data

Algorithm

Outlier candidates

-0.15, -0.07, 20.3, -0.11

Majority

14.51, 17.34, 5.33, 6.01, 0.95, 5.63, 0.98, 19.59, 6.52, 1.63, 0.66, 9.53, 10.16, 12.05, 6.72, 18.47, 17.17, 16.06, 5.24, 6.62, 7.24, 3.57, 17.36, 19.32, 3.79, 1.76, 8.83, 17.06, 11.18, 0.43, 4.76, 13.46, 8.69, 17.47, 11.33, 10.33, 5.63, 17.65, 3.07, 8.37, 15.08, 17.9, 6.53, 4.77, 10.55, 1.52

Outlier

$$Y_t = f(X_t) + \delta_t$$

- Outliers: The observations far away from the fitting function deduced from a subset of the given observations. (Tsaih and Cheng, 2009, page 162)
- Example Rule:

  If $|\delta_t| \geq 3\sigma$, where σ *is* the standard deviation obtained from the model, the t^{th} instance is treated as the outlier.

Keep Vs. Take Out !?

- To purify the data for processing.
  - Data cleansing in data mining → then data modeling
- Outliers will diminish forecast accuracy in time series data. (Chen and Liu, 1993)

- One nurse responded $42,000 as her hourly rate, now that's one **well** paid nurse!
- Averaged wage about $12.00 per hour with a standard deviation of about $2.00.

(Brewer, Nauenberg, & Osborne, 1998)

Outlier detection

- Outlier detection: Task as detecting and removing anomalous observations from data. (Hodge and Austin, 2004)
  - anomaly detection, noise detection, deviation detection, and exception mining
  - In specified domain: intrusion detection, fraud detection, fault detection …
- Outlier Detection Method
  - ➢ Statistical methods
  - ➢ Evolutionary algorithms
  - ➢ Clustering methods
  - ➢ Artificial neural networks

Evolutionary algorithms

- Crawford and Wainwright's research (1995): the best combination is genetic algorithm and Cook's squared distance formula (Cook and Weisberg, 1982).
- Srinoy (2007) implemented a supervised two-phased method to cope with intrusion detection in networking security.
  - This method use **particle swarm optimization** to select the feature for **support vector machine** to classify the intrusion from others. (KDD 99 data set)
- Banerjee's (2012) continuous research has combined genetic algorithm with Euclidean distance due to the feature of density-based distance.

Peer group analysis

- Ferdous and Maeda (2006) implement peer group analysis (PGA) to cope with fraud detection in financial time series data.
  - ➢ Kinds of unsupervised technique featuring its mechanism as identifying peer groups for all the target object

The Following processes are involved in PGA.

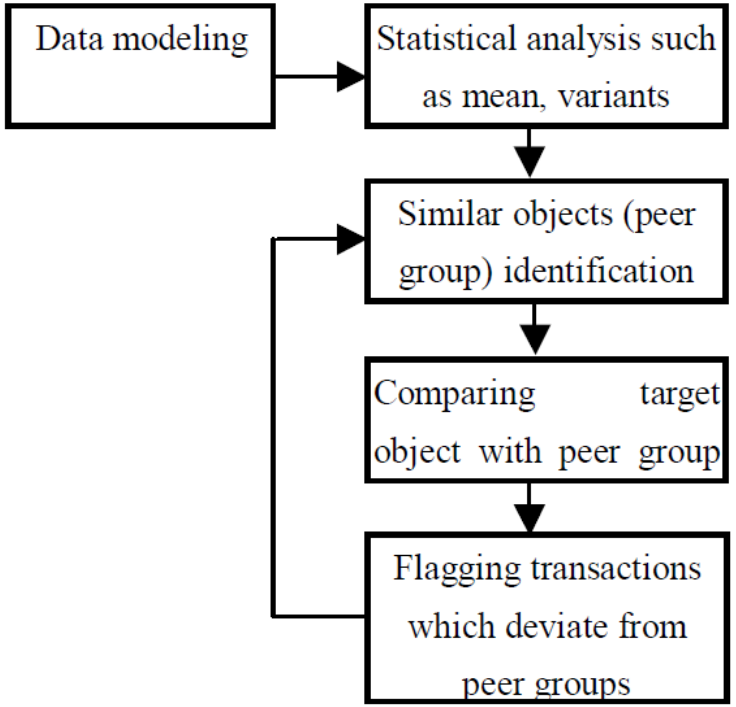


Figure 1: Process Flow of PGA

k-means clustering

- Yoon, Kwon and Bae (2007) tried to use k-means clustering method to detect outliers in software measurement data.
  - The last process of this approach will export an outlier candidate report, this report is still need to be reviews by the domain expert.

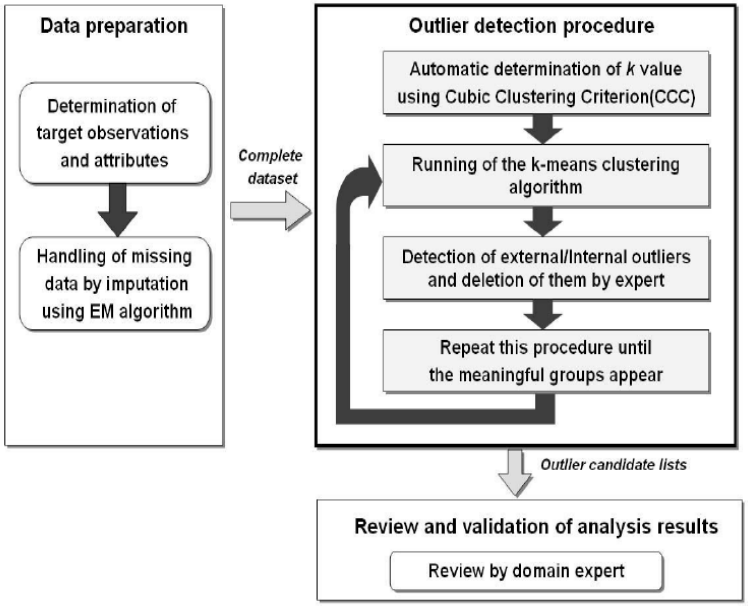


Figure 2. Overall approach.

Artificial neural networks

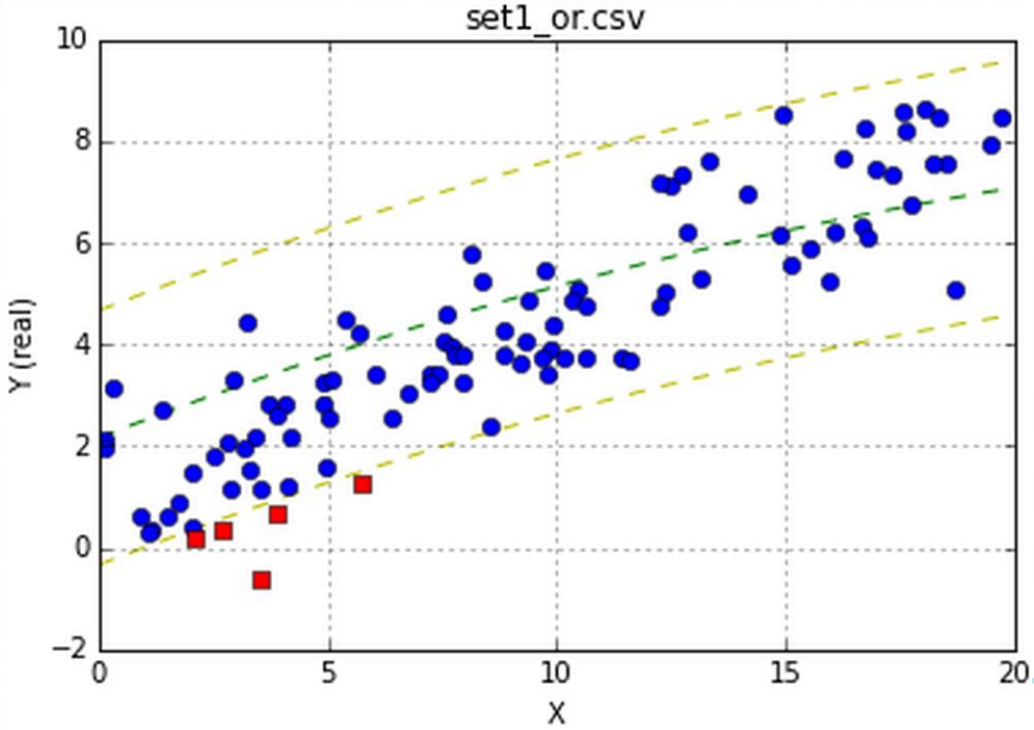
- Sykacek (1997) used a neural network and sigmoid activations to cope with outlier detection problems, where neural network is trained by **Bayesian interface**.
- Williams and Baxter (2002) use **replicator neural networks (RNN)** to measure whether the instance is an outlier or not.
  - The RNN which they proposed is a feed-forward multi-layer perception with three hidden layers between input and output layer.
  - During the RNN is training, the weights of RNN are adjusted to minimize the mean square error.
- The aforementioned works use the pre-specified and fixed network during the training process. They can merely adjust or tuning the weights.

Artificial neural networks

- Tsaih and Cheng (2009) propose that a **resistant learning** outlier detection algorithm with a tiny ε value via SLFN.
  - the **robust procedures** are those whose results are not influenced significantly by violations of the model assumptions (such as when the errors are normally distributed)
  - the **resistant procedures** are those whose numerical results are not influenced significantly by outlying observations
- Huang et al. (2014) detect anomalous pattern effectively in **non-changing** environment. They propose an envelope module to distinguish outliers.

Envelope module

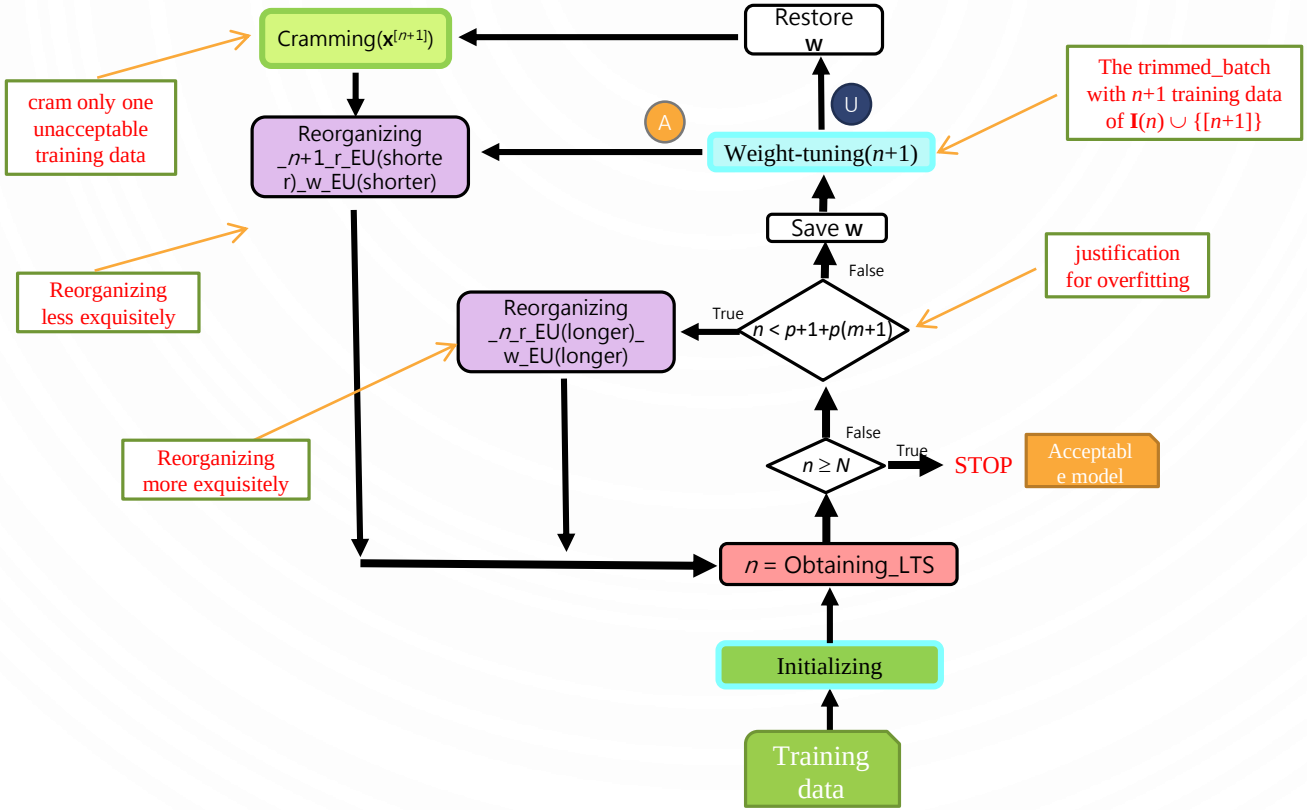
The envelope module allow us to wrap the response elements seen as inliers in the envelope.



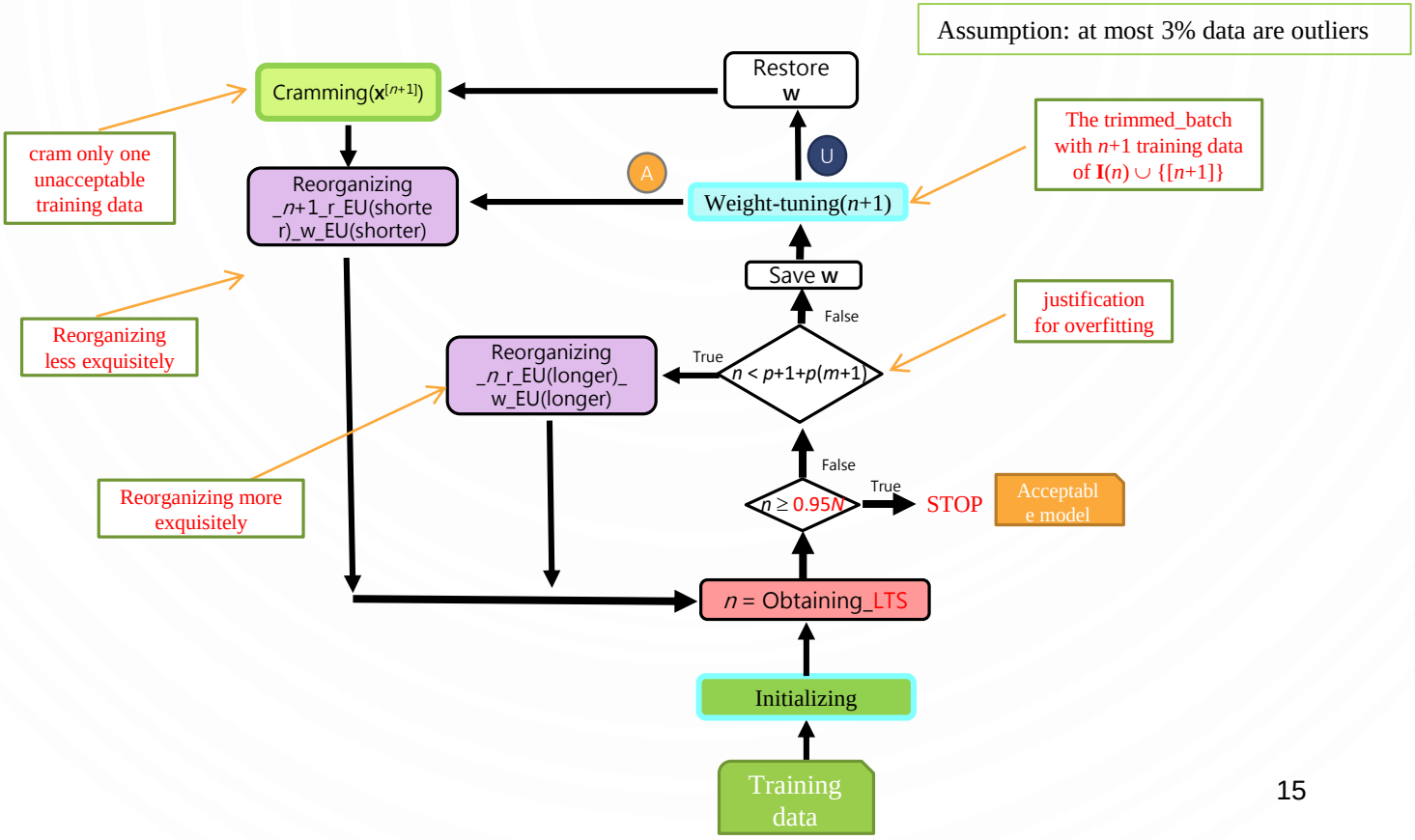
Artificial Neural Networks

- Tsaih, Rua-Huan, Huang, Shin-Ying, Lian, Mao-Ci and Huang, Yennun (2018). “ANN Mechanism for Network Traffic Anomaly Detection in the Concept Drifting Environment,” IEEE DSC 2018, pp. 1-6, 2018.
- Shin-Ying Huang, Jhe-Wei Lin, Rua-Huan Tsaih (2016). “Outlier Detection in the Concept Drifting Environment,” the International Joint Conference on Neural Networks (IJCNN), pp. 31-37.

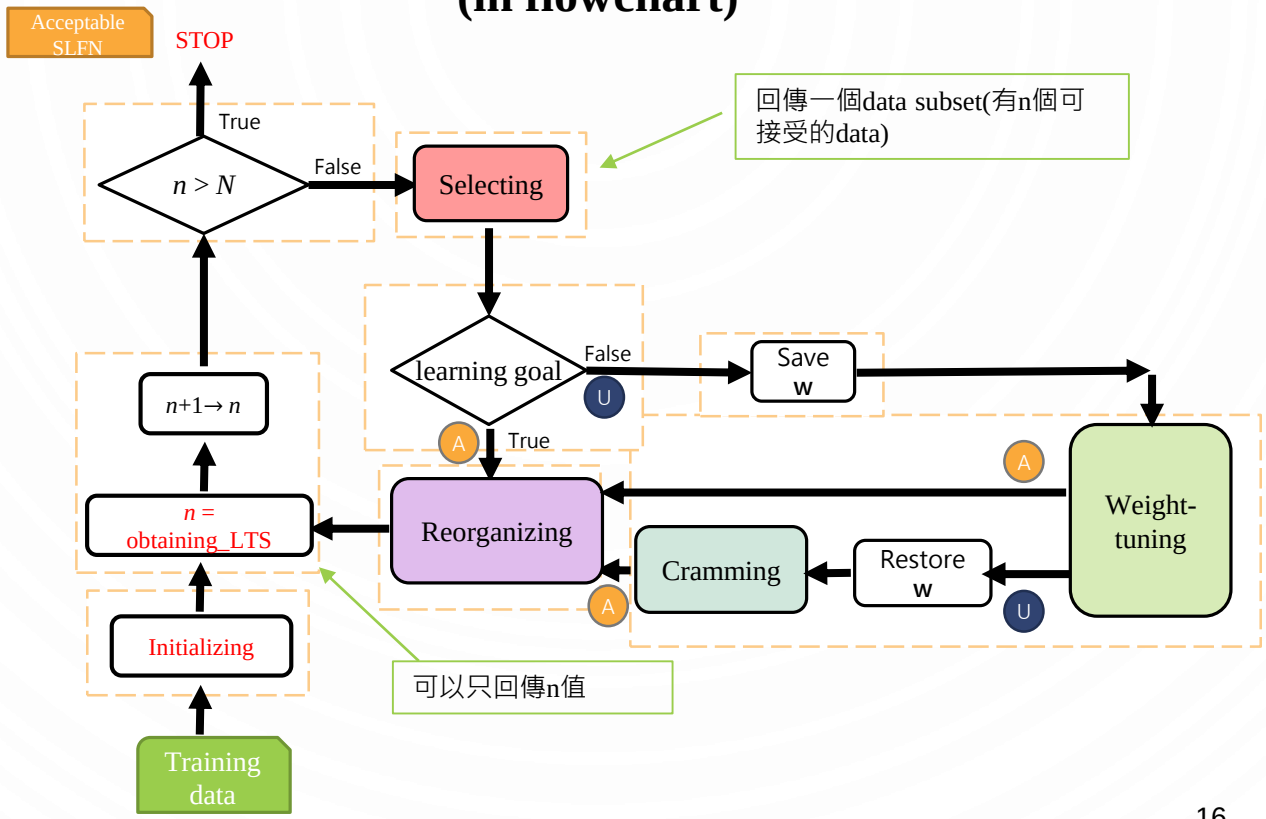
The second new learning mechanism
(in flowchart)



The second new learning mechanism for outlier detection
(in flowchart)



The third new learning mechanism
(in flowchart)



The third new learning mechanism for outlier detection
(in flowchart)