

**P**R**E**D**I**C**T**I**V**E **D**Y**N**A**M**I**X** **I****N****C**

## **Cluster-based Methods for Novelty Detection**

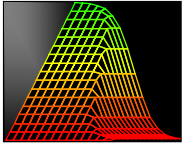
**By Paul Duke,  
Predictive Dynamix, Inc.**

Overview: In addition to being powerful tools for classification and prediction, cluster-based models are also useful in determining whether new data is normal (similar to data used during the training process) or if it is novel (i.e., an outlier). Understanding when new data cases are novel can be extremely important in qualifying the confidence of other predictive modeling methods and identifying behavior in data that has not been previously encountered. Applications exist in fraud detection, process drift/fault detection, homeland security, preventative maintenance, model lifecycle management, and many other areas.

### Supervised vs. Unsupervised Learning

Forecasting and classification methods are extremely useful for identifying known patterns and predicting outcomes by using historical data. Least squares, backpropagation, and C4.5 are “supervised” methods used for training regression, neural network, and decision tree models, respectively. They are considered supervised in that they learn to approximate known output target values by adjusting model parameters (coefficients, rules, etc.) in order to minimize the model’s error (mean squared error, classification error, etc.).

Cluster-based models can be trained in a supervised manner with the purpose of generating a prediction or classification. However, they are also useful in characterizing the distribution of data without training to generate a specific target result. Learning algorithms that train with no knowledge of a specified target output are considered “unsupervised” learning methods.



# PREDICTIVE DYNAMIX INC

## Cluster-based Models

Core to the training process of cluster-based methods are algorithms that group similar cases together into clusters (a.k.a. segments). “Similarity” is related to “distance” in the sense that the greater the similarity between two data points, the lesser the distance is between them. Typically, the distance metric used is classical, Euclidean distance ( $d = \sqrt{\sum (X_i - Y_i)^2}$ ). So, the distance between two data points (or a data point and a cluster’s centroid) is easily computed as the Euclidean distance between the two data points.

Cluster training methods distribute a set of clusters across the data set, effectively assigning similar cases to a given cluster. At runtime, a trained cluster-based model calculates the distance between the incoming data point and each of the fixed (trained) clusters. Whichever cluster is closest to the data point is said to be the “winning” cluster.

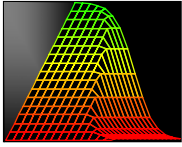
## Novelty Detection via Cluster Distance Statistics

It can be the case that, even when a new data point is assigned to a winning cluster, the distance from the data point to winning cluster may be further than normal. In the instance when the new data point is further away from the winning cluster than the maximum training distance, the data point is considered novel.

Another approach is to calculate mean and standard deviation statistics for each cluster’s training distances. Then, establish a novelty limit where a new data point’s distance from its winning cluster exceeds some number of standard deviations from the mean (training) distance for that cluster.

## Novelty Detection via Cluster Variable Statistics

A more stringent method for identifying novel data is to compare the variable values of the data point to the max and min variable values that were encountered during training – for each cluster. During runtime, when a new data point is assigned to a cluster, each



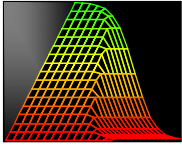
# PREDICTIVE DYNAMIX INC

variable is then checked against the max and min variable threshold limits for the cluster. If any variable value is beyond the max/min limits, the data point is considered novel.

A similar approach is to calculate mean and standard deviation statistics for each cluster's training variable values. Then, establish a novelty limit when any of a new data point's values exceed some number of standard deviations from the mean for each variable.

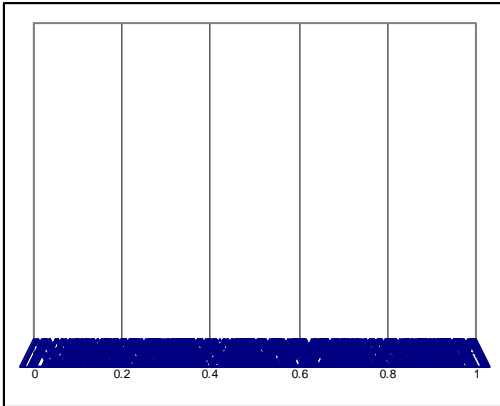
## Novelty Detection via Cluster Win Frequency

In some cases, all data won by a cluster may be determined to be novel. This may be the case if the cluster had a low training win frequency (very few cases assigned to it) or because of specific knowledge of the problem domain. In this event, any new data points that map to this cluster are considered novel.



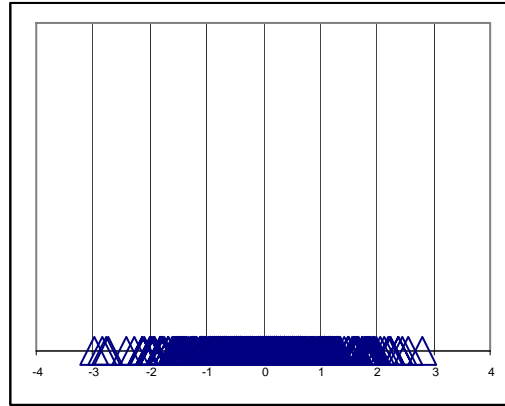
## Univariate Examples

In situations where a “data point” consists of a single variable, outlier detection is often done by using simple, univariate statistics (Max, Min, Mean, Stdev, etc.) as threshold rule limits for detecting novel cases.



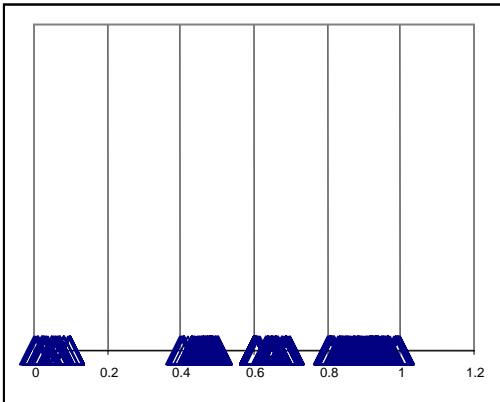
For a uniform data distribution, using max & min values may be sufficient to detect novel values.

IF (  $x > \text{MaxXValue}$  ) OR (  $x < \text{MinXValue}$  )  
THEN DataIsNovel = True

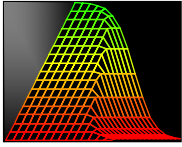


For a normal (Gaussian) distribution, a limit on the number of stdev's from the mean may be effective.

IF ABS( (  $x - \text{MeanX}$  ) / StdevX ) > 3.5  
THEN DataIsNovel = True



In cases where the data is neither uniformly nor normally distributed, even single variable models can benefit from using clustering for novelty detection because no single set of variable statistics adequately describes the data. Writing manual rules to describe novelty limits can get quite involved very quickly.



## Multivariate Example

In real world applications, it is often the case that a given variable's distribution may be interrelated with one or more other variables. In these situations univariate statistics fall far short as effective novelty detectors. As shown below, data values can easily fall within their normal variable range limits and still be quite novel.

As the number of model variables increases, cluster-based models are ideal automatic novelty detectors because they excel at characterizing multivariate data distributions.

