Outlier Analysis, Research Frontiers Assignment

# Outlier Analysis

There are three types of outliers: global, contextual, and collective. A global outlier may be something like a patient having a much higher white blood cell count than other patients. This can be detected with a clustering-based method such as density-based proximity clustering. A contextual outlier can be exemplified by a lower temperature in summer time; think 80 degrees on an Augustus day in Texas. 80 degrees is not an abnormal temperature to have in Texas, but it is almost unheard of in the month of August since temperatures are usually 90 degrees and above. This outlier can be detected by first reducing the dataset to summer months and then using statistical models to create clusters. Finally, a collective outlier could be something like a group of whales moving slower on their migration path. A majority/minority method of clustering can be used to determine which whale pods are the slowest within their species. This is accomplished by taking the whale pod as one data point rather than each individual whale.

# Research Frontiers: Semi-Supervised Learning

Semi-supervised learning is a combination of supervised and unsupervised learning, or classification and clustering. It is used in settings where classifications exist but it is difficult to label all the data points. This especially important in domains where niche knowledge is needed to add labels but human resources are limited. SSL is made difficult due to its disjointed use of two different methods but is much more accurate than one of these methods alone, particularly for clusters that do not take on a circular form.

Ways of solving this hybrid issue include transductive learning, active learning, and inductive learning. In transductive learning, the algorithm takes known labeled points and propogates labels to nearby points. Active learning is used to predict which unlabeled points would be most useful in predicting the rest of the dataset’s labels. This is different from transductive learning in that it does not automate the labeling itself. Finally, inductive learning trains a classification model with all points in the dataset. Once a model has been created and all unlabeled data given pseudo-labels, the model can be retrained based on how accurate it was. This is called the wrapper method. When applying a probabilistic model to this instead of a deterministic model, it is called self-training.