Data Stream Mining- Lecture 9

Novelty Detection in Data Streams

Chandresh Kumar Maurya, Research Assistant Professor

Eötvös Loránd University, Budapest, Hungary

November 19, 2019

Novelty Detection

Definition

Novelty detection refers to the identification of new concepts, change in the old concepts or the presence of noise.

Synonyms terms:

- Outlier detection
- One-class classification
- Anomaly detection

Definitions

Definition (Novelty)

Novelty is a concept represented by a group of examples sharing some characteristics.

Definition (Outlier)

A sparse, independent examples whose characteristics different from the normal examples are called outliers

Definition (Anomaly)

A novel concept which is unexpected, abnormal in a specific domain or application such fault detection, spam classifcation etc.

Desiderata for Novelty Detection

1. **Principle of robustness:** A novelty detection method must be capable of robust performance on test data that maximizes the exclusion of novel samples while minimizing the exclusion of known samples.
2. **Principle of generalization:** The system should be able to generalize without confusing generalized information as novel
3. **Principle of independence:** The novelty detection method should be independent of the number of features, and classes available.
4. **Principle of adaptability:**A system that recognizes novel samples during test should be able to use this information for learning new concepts
5. **Principle of computational complexity:** A number of novelty detection applications are online and, therefore, the computational complexity of a novelty detection mechanism should be as low as possible.

Basic Framework

- Getting labelled data is problematic (why?)

Basic Framework

- Getting labelled data is problematic (why?)
  - Don't have expert.
  - Not sufficient time in the case of streaming data.
  - Sometimes, labeling is expensive
- Build model in *Offline* phase using a small amount of labeled data.
- Use model to predict the new data point

But, how does the *novelty and concept-drifts* are handeled?

Basic Framework

- Getting labelled data is problematic (why?)
  - Don't have expert.
  - Not sufficient time in the case of streaming data.
  - Sometimes, labeling is expensive
- Build model in *Offline* phase using a small amount of labeled data.
- Use model to predict the new data point

But, how does the *novelty and concept-drifts* are handeled?

1. First create micro-clusters using initial-set of labeled data
2. Then, for each incoming data point, calculate its distance from the centroid of the micro-clusters. If the distance is more than a user-specified threshold, put in a buffer and after a enough number of points, declare as novelty.

Online Clustering for Novelty Detection and Concept Drift in Data Streams [Garcia et al., 2019]

The algorithm proposed in the above is called Higia.

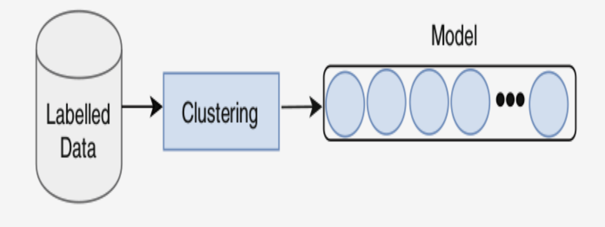

Figure: Higia Offline

Contd...

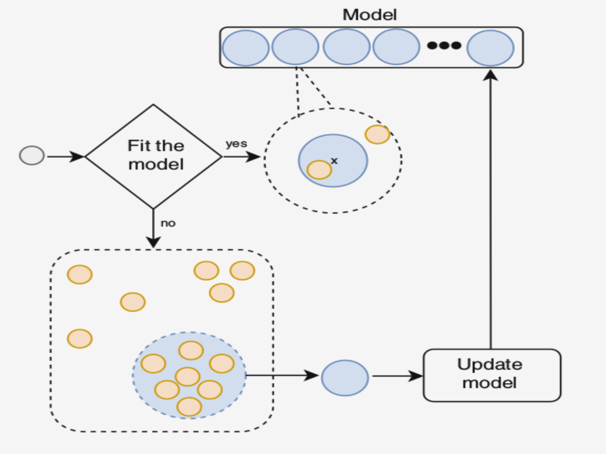

Figure: Higia Onine

Contd...

Algorithm 1. Higia: Online Phase

```
input: X_tr, T, k
Let ψ_k be a list of the k nearest micro-clusters to X_tr
if majority of ψ_k have the same label then
    Let C_j be the nearest micro-cluster to X_tr
    Let c_j be the centroid of C_j
    Let radius(C_j) be the radius of C_j
    dist ← EuclidianDistance(X_tr, C_j)
    if dist ≤ radius(C_j) then
        update C_j with X_tr
        classify X_tr with the same label of C_j
    else if dist ≤ (radius(C_j) × T) then
        create extension of C_j with centroid X_tr and radius 0.5
        classify X_tr with the same label of C_j
else
    add X_tr to buffer
    classify X_tr as unknown
```

Some results

Statistics	1CDT	MOA	Gear	UG	SynD	Forest Cover
Attributes	2	4	2	2	10	54
Classes	2	4	2	2	2	7
Normal classes	1	2	2	1	2	3
New classes	1	2	0	1	0	4
Instances MinCla	7199	9987	99935	44999	124660	587
Instances MajCla	7200	18180	100065	45000	125340	18350

Figure: Data set

Some results

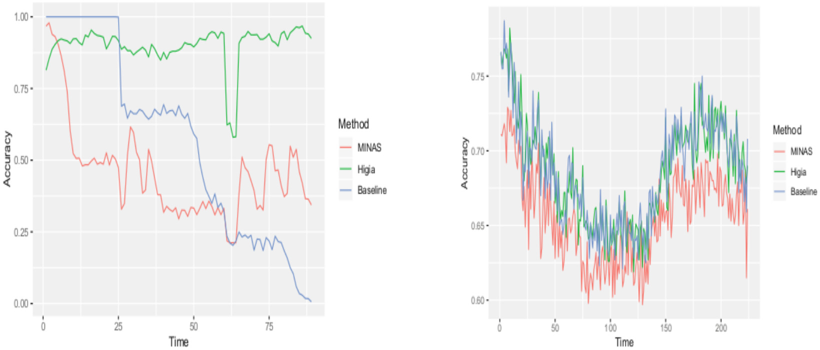

Figure: Accuracy over time

Some results

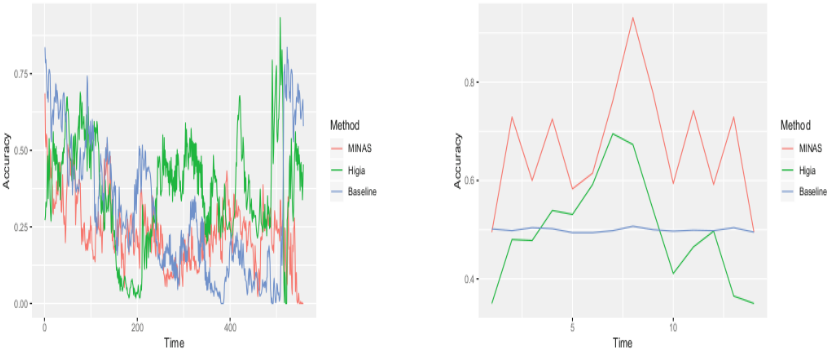

Figure: Accuracy over time

Bibliography I

Garcia, K. D., Poel, M., Kok, J. N., and de Carvalho, A. C. (2019).
Online clustering for novelty detection and concept drift in data streams.
In *EPIA Conference on Artificial Intelligence*, pages 448–459. Springer.