电子科技大学
University of Electronic Science and Technology of China

Concept Drift

Ke Yan

Data Mining Lab, Big Data Research Center, UESTC
Email：junmshao@uestc.edu.cn
http://staff.uestc.edu.cn/shaojunming

Outline

1. What is concept drift
2. Concept drift Classification
3. Concept drift detector Classification
4. Concept drift detect methods
5. Experiment
6. Open issue

What is concept drift

In predictive analytics and machine learning, the concept drift means that the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways.

----------Wikipedia

In a word, the probability distribution changes.

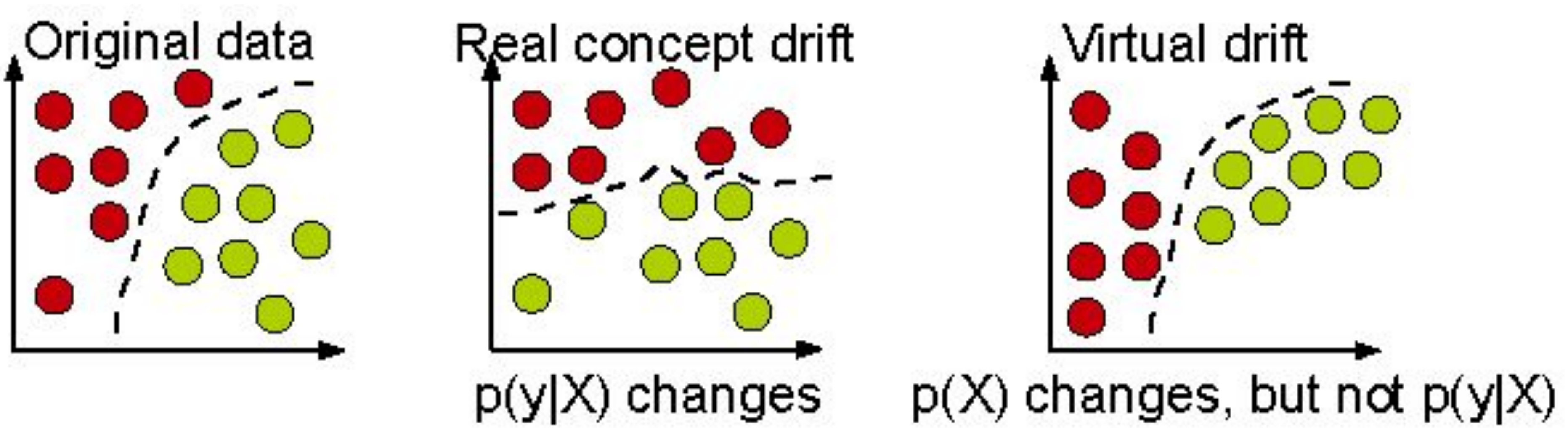
Potential causes:

- Change in P(C)
- Change in P(X|C)
- Change in P(C|X)-----Classification

Concept drift Classification

Reason of change

Real concept drift *VS* Virtual concept drift



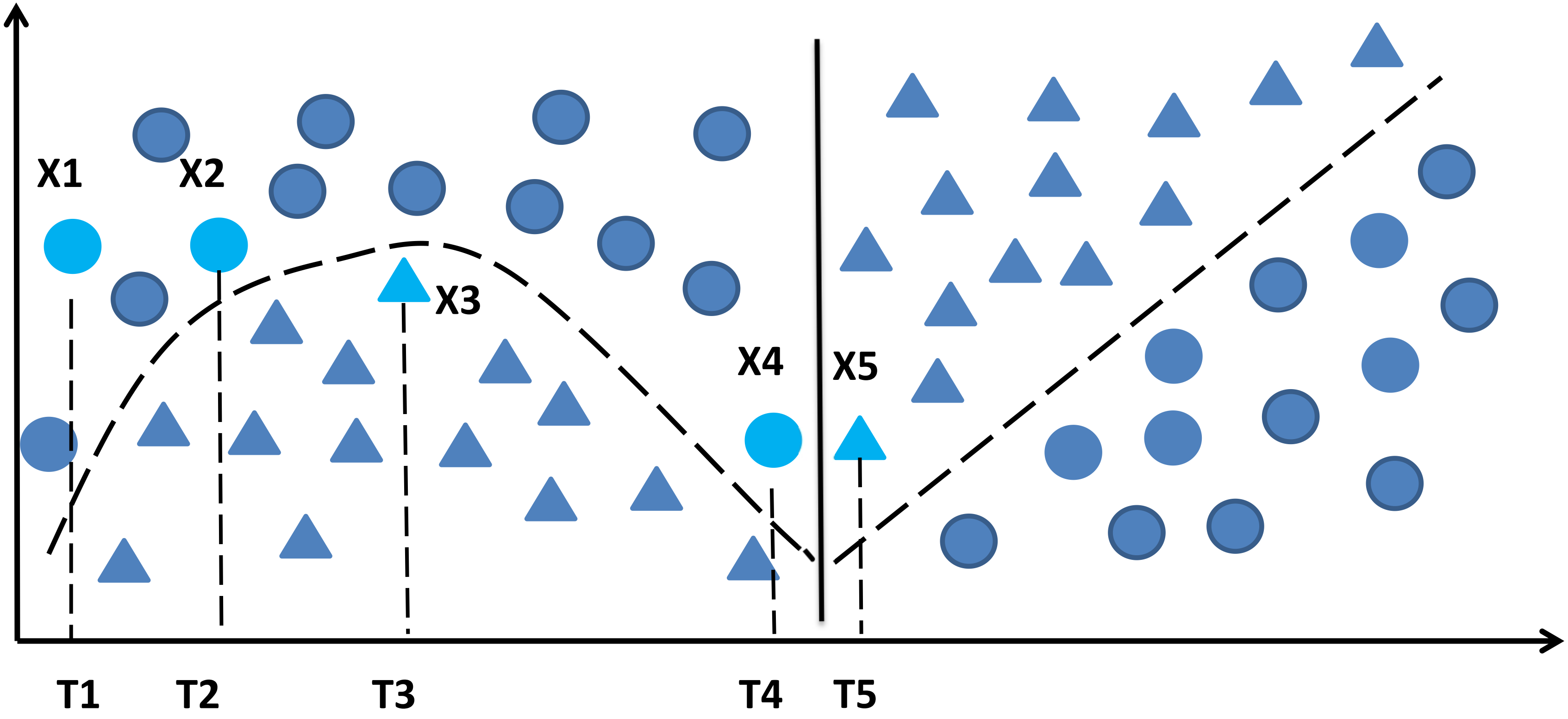
(Gama et al., 2014)

$$P(C_i \mid X) = \frac{P(C_i)\,P(X \mid C_i)}{P(X)}$$

Concept drift Classification

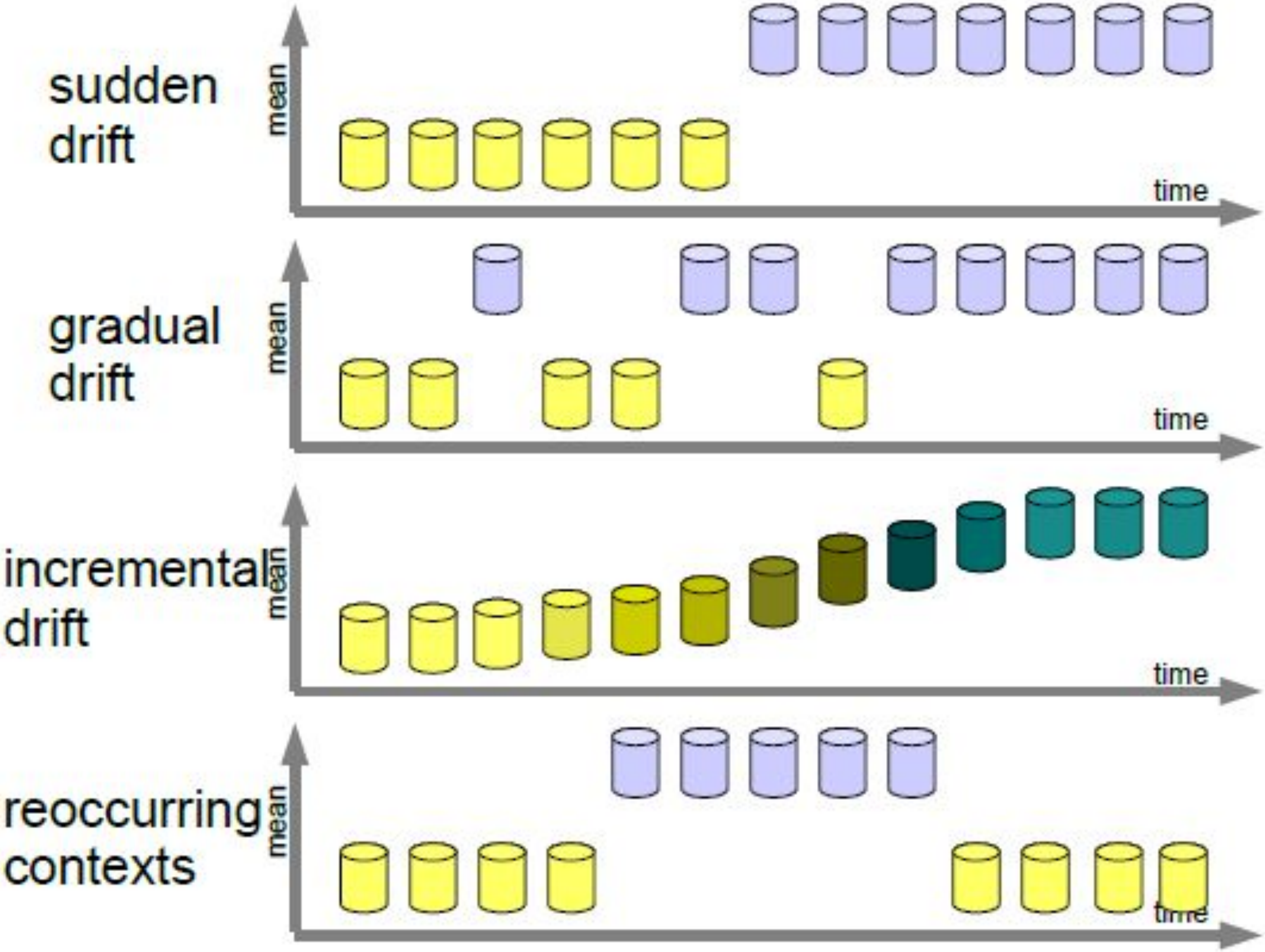
Speed of change

Gradual concept drift *VS* Abrupt concept drift



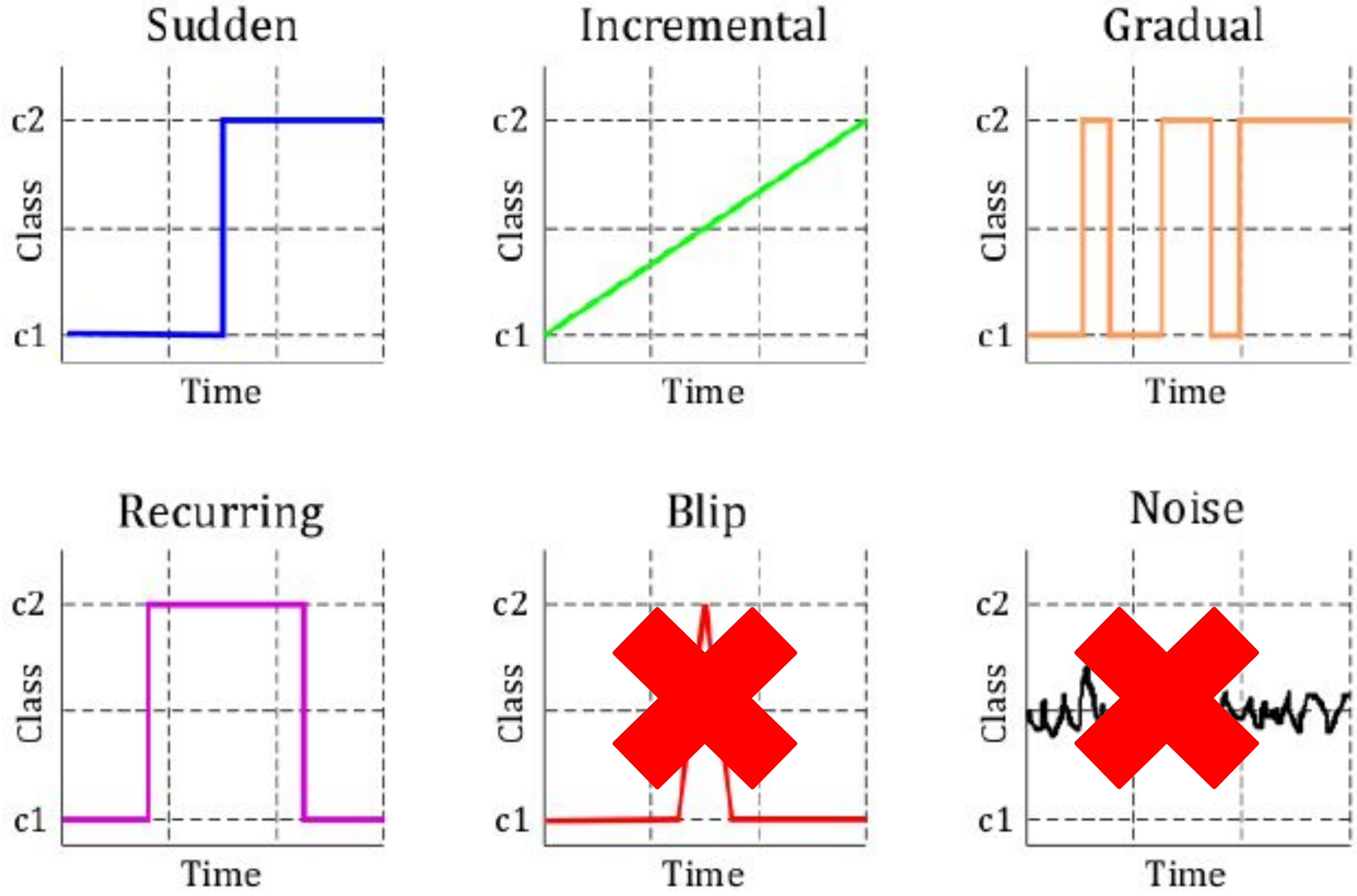
Concept drift Classification

Types



Concept drift Classification

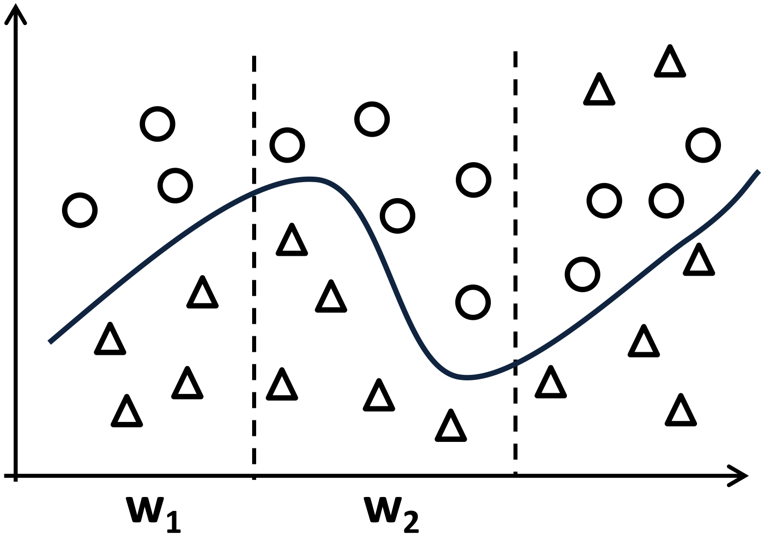
Types



Concept drift detector Classification

Distribution-based detector

Monitoring the change of data distributions between two fixed or adaptive windows of data. (e.g. ADWIN)

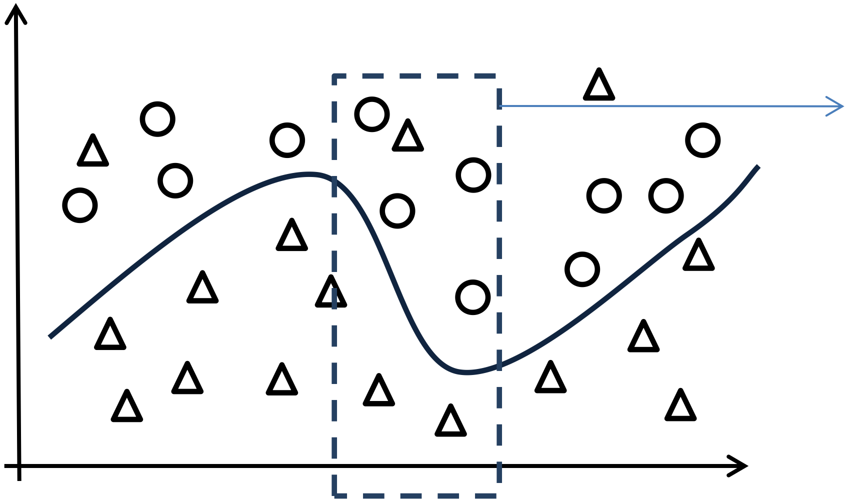


- Hard to determine window size
- Learn concept drift slower
- Virtual concept drift

Concept drift detector Classification

Error-rate based detector

Capture concept drift based on the change of the classification performance. (i.e. comparing the current classification performance to the average historical error rate with statistical analysis.) (e.g. PHT)



Infer time-changing concept relying on **performance degradation**

Indirectly

- Sensitive to noise
- Hard to deal with gradual concept drift
- Depend on learning model itself heavily

Concept drift detect methods

Single Classifier

Learn the new instance by the basic learner, if correctly classify, send the value to the concept drift detector to testing whether drift, level of alert level or drift.

- DDM(Error-rate based)
- EDDM(Error-rate distance based)

PL(Paired Learners) : a stable learner and a reactive learner. The stable learner based on all of its experience, the reactive one predicts based on a window of recent examples.

ADWIN(Adaptive Windowing): monitor the change rate of data in the slid window. Stable increase the window, otherwise decrease. Two sub window, compare the data distribution. Chi-square test.

Concept drift detect methods

STEPD (Statistical Test of Equal Proportions): compare the accuracy of recent w instance with the all distance.

ECDD(Exponentially Concept Drift Detection): Exponentially Weighted Moving Average (EWMA) , with weight ***VS*** without.

PHT (The Page-Hinkley Test): the observed values ***VS*** the mean. Concept drift, then the actual accuracy decrease.

DOF(The Degree of Drift): find a nearest neighbor in the previous chunk, compare the label. Bulit the distance map(index to label). Drift degree mean ***VS*** standard.

Concept drift detect methods

Performance Comparison

1. SingleClassifier with DDM or EDDM good performance in a stable environment and compute fast.
2. PL good performance in a unstable environment， the worst in a stable environment。 (High false alarm)
3. HoeffdingAdaTree is based on the Hoeffding tree， inconsistency of the performance 。
4. Ensemble learning approaches, WeightedEnsemble, OzaBagAdwin and PASC can adapt to the concept drift, but can’t capture the concept drift, so the predict performance is worse。

Experiment---Datasets

Synthetic dataset

- Extreme verification latency benchmark
- Sine, Line, Plane, Circle and Boolean Data Sets
- SEA concepts
- STAGGER

Real-word dataset

- Airline approximately 116 million flight arrival and departure records
- Chess.com (online games) and Luxembourg (social survey).
- Sensor stream and Power supply stream datasets
- ECUE spam consisting of more than 10,000 emails
- Elec2 2 classes, 45312 instances
- Text mining a collection of text mining datasets
- Gas Sensor Array Drift Dataset Data Set
- PAKDD'09 competition credit evaluation
- KDD'99 competition simulated intrusions in a military network environment

More Dataset Click: http://www.liaad.up.pt/kdus/products/datasets-for-concept-drift

Experiment--- Evaluation Metrics

False Alarm Rate(FA)

$$FD = \frac{error\_alarm}{all\_alarm} \times 100\%$$

Miss Detection Rate(MD)

$$MD = \frac{miss\_alarm}{all\_alarm} \times 100\%$$

Confirmation Delay(CD)

$$CD = InstanceNumberOf\,(Confiromdrift - \mathrm{Re}\,aldrift)$$

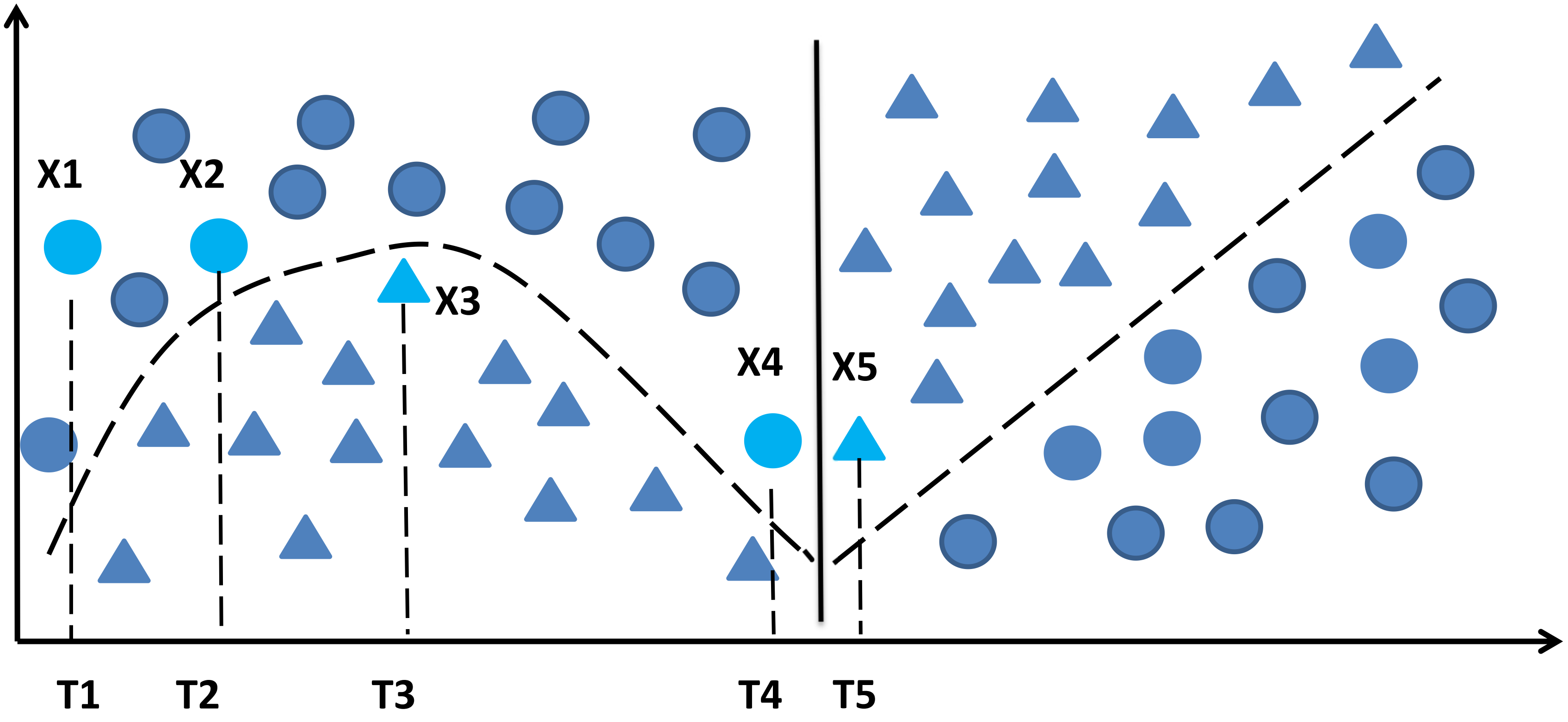
Warning Delay(WD)

$$WD = InstanceNumberOf\,(Warmdrif - \mathrm{Re}\,aldrift)$$

Number Of Drift Detected(DD)

Sensitivity *VS* Robustness

Experiment--- Evaluation Metrics



$$CD = InstanceNumberOf(T3 - T1)$$

$$CD = InstanceNumberOf(T5 - T4)$$

$$WD = InstanceNumberOf(T2 - T1)$$

$$WD = InstanceNumberOf(T5 - T4)$$

Open Issue

1. Unbalanced distribution of categories
2. Repetitive learning(recurring concept drift)
3. Semi-supervised learning and Active learning
4. Cold Start(Misclassification)
5. Learning single-concept only

Thanks

Ke Yan

keyan2048@gmail.com