

A Feasibility Study of CONCEPT DRIFT in Empirical Software Engineering

Md Alamgir Kabir

Supervisor: Dr. Jacky Keung

October 21, 2018

Department of Computer Science

1. CONCEPT DRIFT
2. Types of Concept Drift
3. Detecting Concept Drift
4. Practical Problem in Classification
5. Implications in Software Engineering Research

CONCEPT DRIFT

Concept Drift

- **Streaming data:** Incoming data of heterogeneous sources
- Behavior of these streaming data is changing over time
- **CONCEPT DRIFT**¹: The changes in the distribution or concept over time
- A time changing probability distribution in high speed data streams
- Concepts are often not stable but change with time such as Weather Prediction and Customers' Preference

¹Ditzler, G. (2015). Learning in Nonstationary Environments: A Survey, (November). <https://doi.org/10.1109/MCI.2015.2471196>

Concept Drift

According to Lu ², the problem can be framed as follows.

- If we denote *the feature vector as x* and the *class label as y* , then the *data stream* will be an infinite sequence of (x, y) .
- If the concept drifts, it means the distribution of $p(x, y)$ is changing between the current data chunk
- If we decompose $p(x, y)$ into the following two parts as $p(x, y) = p(x) \times p(y|x)$
- we could say there are two sources of concept drift: *one is $p(x)$* , which evolves with time t , and can also be written as $p(x|t)$
- the *other is $p(y|x)$* , the conditional probability of feature x

²Lu, N., Zhang, G., & Lu, J. (2014). Concept drift detection via competence models. Artificial Intelligence, 209(1), 11–28.

<https://doi.org/10.1016/j.artint.2014.01.001>

Types of Concept Drift

Types of Concept Drift

According to Žliobaite, ³

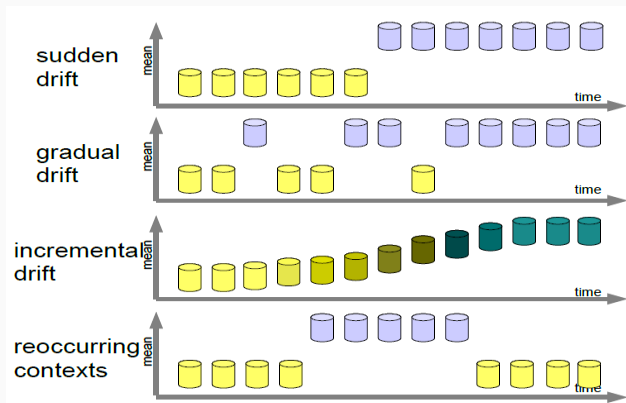


Figure 1: The four structural types of the drift

³Žliobaite, I. (2010). Learning under concept drift: an overview. arXiv preprint arXiv:1010.4784.

Detecting Concept Drift

Detecting Concept Drift

The first attempt to handle of concept drift with case-based technique was IB3 (Instance-based) learning ⁴.

Detecting concept drift-

- Monitoring **raw data** (By data distribution)
- Monitoring **parameters of learners** (By parameters)
- Monitoring prediction **errors of learners** (By learners output)

⁴Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine learning, 6(1), 37-66.

Practical Problem in Classification

Practical problem in classification

Being the environments non stationary it exhibits class imbalance that causes the practical problem in classification i.e., one of the most major problem in data mining and machine learning ⁵.

- It requires a lot of effort to observe the new minority class instances in new incoming data.
- The techniques of incremental learning in imbalance datasets (i.e., limited label data and plenty unlabeled data) with semi-supervised learning are more complex compare to the techniques of fixed distributions.

⁵Hoens, T. R., & Chawla, N. V. (2012). Learning in non-stationary environments with class imbalance. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12. <https://doi.org/10.1145/2339530.2339558>

Implications in Software Engineering Research

- Prediction model are developed by using historical datasets (i.e., fixed distributions)
- Proposed models to predict the number and the location of future bugs in software source codes
- Such predictions can help the project manager to lead the project by proper utilizing the testing resources
- Unfortunately, the prediction model may not work if the historical datasets changes overtime ⁶

⁶Ekanayake, J., Tappolet, J., Gall, H. C., & Bernstein, A. (2009). Tracking concept drift of software projects using defect prediction quality. Proceedings of the 2009 6th IEEE International Working Conference on Mining Software Repositories, MSR 2009, 51–60.

- Due to the change of some influencing features, bug generation process becomes more unsuitable.⁷
- They observed that the change in number of author editing a file and a number of defects fixed by them introduce the concept drift.
- They have also seen that the prediction quality significantly varies over time.

⁷Ekanayake, J., Tappolet, J., Gall, H. C., & Bernstein, A. (2012). Time variance and defect prediction in software projects. *Empirical Software Engineering*, 17(4–5), 348–389.

Implications in SE Research/3

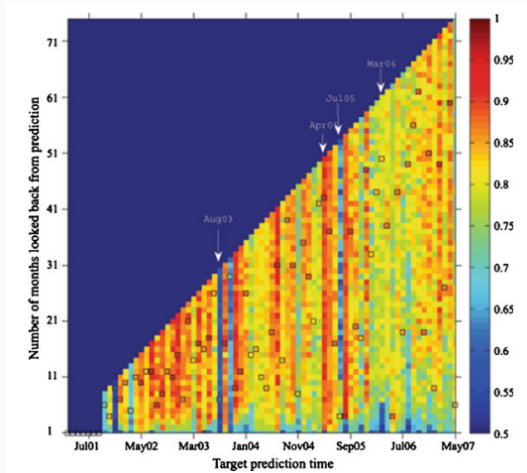


Figure 2: Eclipse heat-map: Prediction quality using different training periods with the points of highest AUC highlighted

- The maximum **AUC values** for each target period typically lie on neither ends.
- The models should not be trained on data collected from a very long or very short history.⁸

⁸Ekanayake, J., Tappolet, J., Gall, H. C., & Bernstein, A. (2012). Time variance and defect prediction in software projects. Empirical Software Engineering, 17(4–5), 348–389.

Current studies demand...

- Current studies demand to do the several experiments in the area of empirical software engineering to prepare the *well trained prediction model in streaming datasets*.
- We have planned a set of experiments in the datasets of *open source projects*.

Thank you so much for listening!