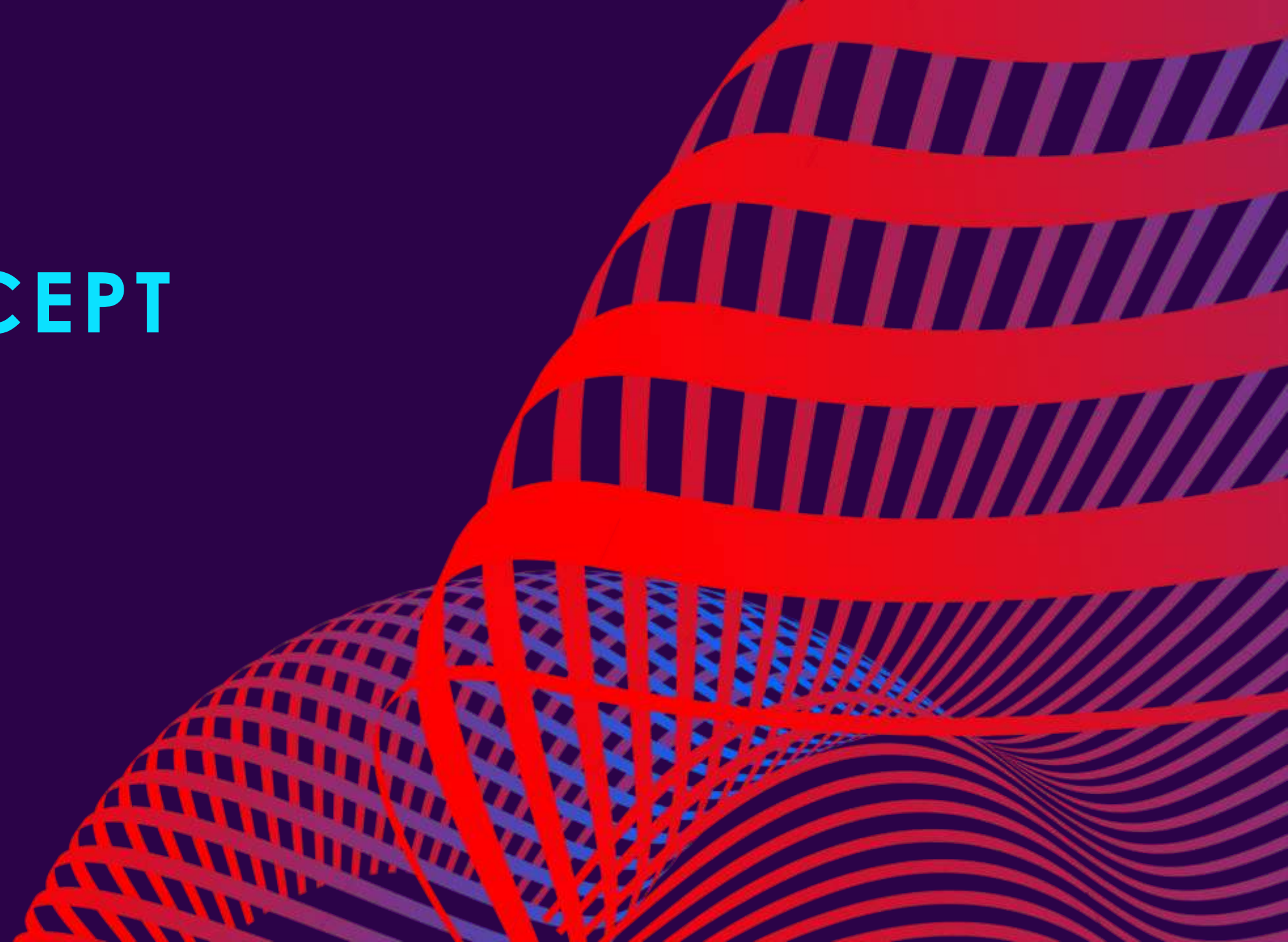
CONCEPT DRIFT

Arda Erol

Osman İlge Ünaldı



INTRODUCTION

Concept drift means that the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways.

PROBLEM DESCRIPTION

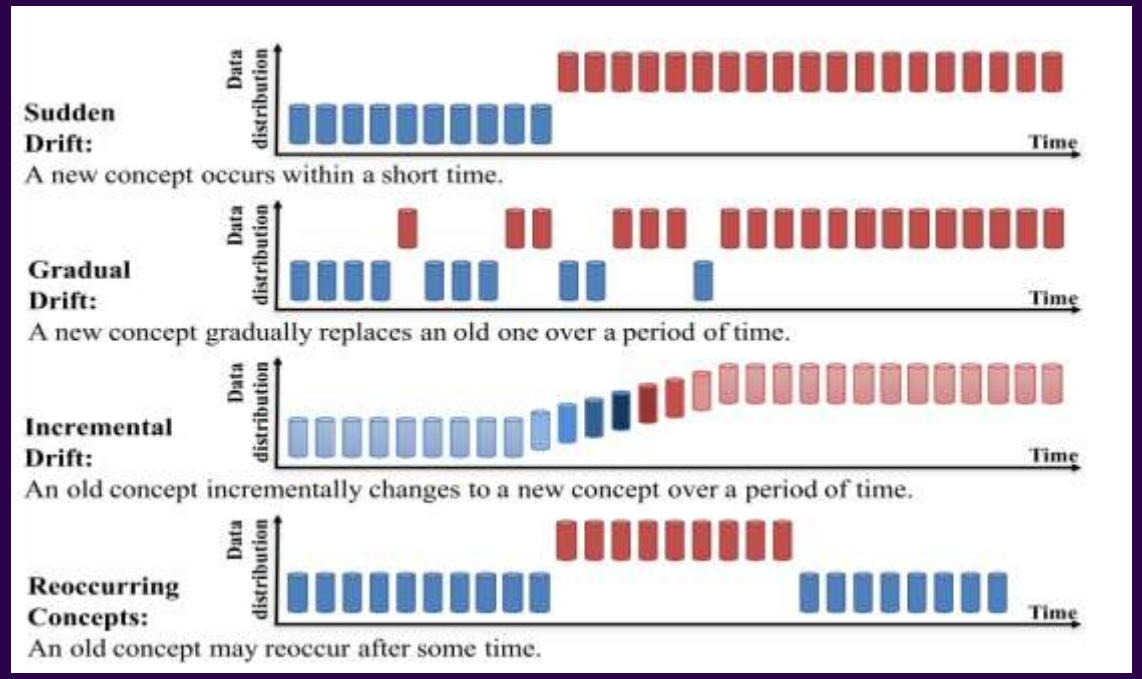
Concept drift is a phenomenon in which the statistical properties of a target domain change over time in an arbitrary way. These changes might be caused by changes in hidden variables which cannot be measured directly.

CONCEPT DRIFT DEFINITION AND THE SOURCES

Source I: $P_t(X) \neq P_{(t+1)}(X)$ while $P_t(y|X) = P_{(t+1)}(y|X)$, that is, the research focus is the drift in $P_t(X)$ while $P_t(y|X)$ remains unchanged. Since PtðXÞ drift does not affect the decision boundary, it has also been considered as virtual drift [7], Fig. 3a. Source II: $P_t(y|X) \neq P_{(t+1)}(y|X)$ while $P_t(X) = P_{(t+1)}(X)$ while $P_t(X)$ remains unchanged. This drift will cause decision boundary change and lead to learning accuracy decreasing, which is also called actual drift, Fig. 3b

Source III: mixture of Source I and Source II, namely $P_t(X) \neq P_{(t+1)}(X)$ and $P_t(y|X) \neq P_{(t+1)}(y|X)$. Concept drift focus on the drift of both $P_t(X)$ and $P_t(y|X)$, since both changes convey important information about learning environment

THE TYPES OF CONCEPT DRIFT



Sudden Drift

Gradual Drift

Incremental Drift

Reoccuring Concepts

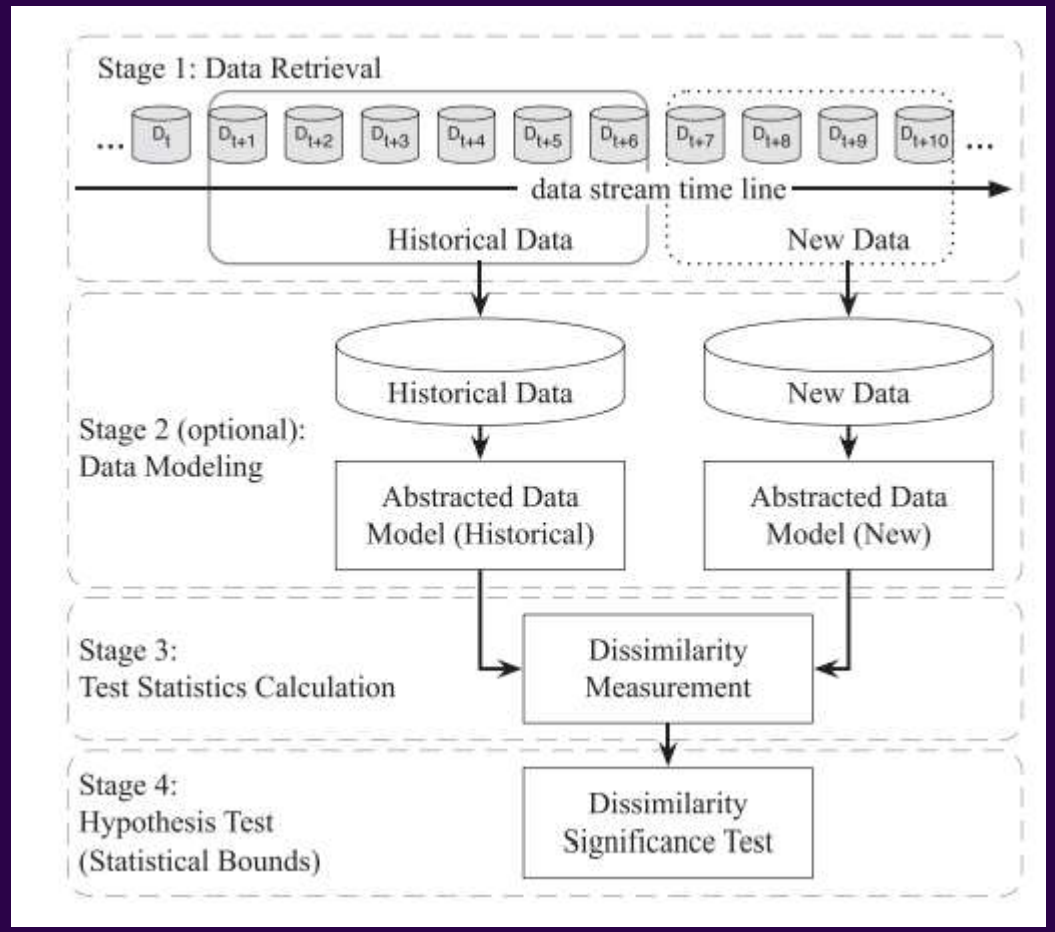
A GENERAL FRAMEWORK FOR DRIFT DETECTION

Stage 1 (Data Retrieval) aims to retrieve data chunks from data streams.

Stage 2 (Data Modeling) aims to abstract the retrieved data and extract the key features containing sensitive information, that is, the features of the data that most impact a system if they drift.

Stage 3 (Test Statistics Calculation) is the measurement of dissimilarity, or distance estimation. It quantifies the severity of the drift and forms test statistics for the hypothesis test.

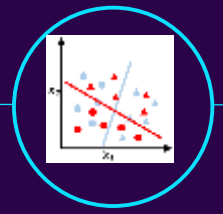
Stage 4 (Hypothesis Test) uses a specific hypothesis test to evaluate the statistical significance of the change observed in Stage 3, or the p-value.



CONCEPT DRIFT DETECTION ALGORITHMS

Error Rate

These algorithms focus on tracking changes in the online error rate of base classifiers.



Data Distribution

Algorithms of this category use a distance function/metric to quantify the dissimilarity between the distribution of historical data and the new data. If the dissimilarity is proven to be statistically significantly different, the system will trigger a learning model upgradation process.

Multiple Hypothesis

The novelty of these algorithms is that they use multiple hypothesis tests to detect concept drift in different ways. These algorithms can be divided into two groups: 1) parallel multiple hypothesis tests; and 2) hierarchical multiple hypothesis tests.

CONCEPT DRIFT UNDERSTANDING

Drift understanding refers to retrieving concept drift information about “When”, “How”, and “Where”. This status information is the output of the drift detection algorithms and is used as input for drift adaptation.

HOW

The severity /degree of concept drift.

WHEN

The time at which the concept drift occurs and how long lasts.

WHERE

The drift regions of concept drift

DRIFT ADAPTATION

TRAINING NEW MODELS FOR GLOBAL DRIFT

If the stable learner frequently misclassifies instances that the reactive learner correctly classifies, a new concept is detected, and the stable learner will be replaced with the reactive learner.

MODEL ENSEMBLE FOR RECURRING DRIFT

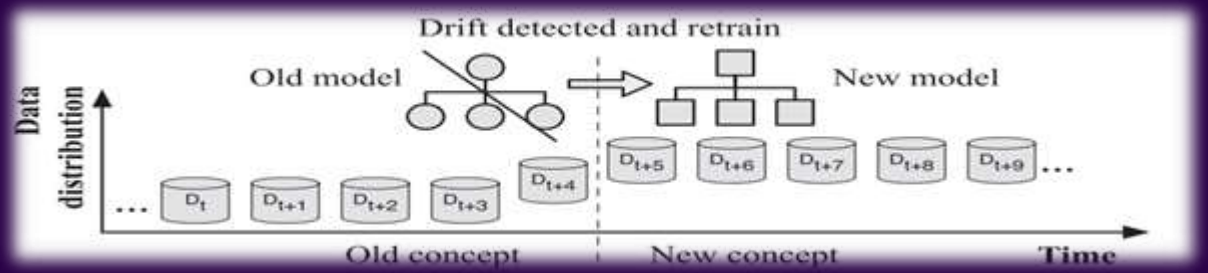
In the case of recurring concept drift, preserving and reusing old models can save significant effort to retrain a new model for recurring concepts.

ADJUSTING EXISTING MODELS FOR REGIONAL DRIFT

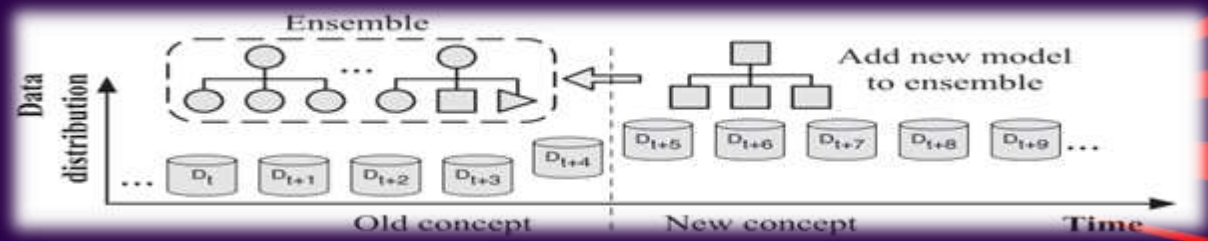
An alternative to retraining an entire model is to develop a model that adaptively learns from the changing data. Such models have the ability to partially update themselves when the underlying data distribution changes.

DRIFT ADAPTATION

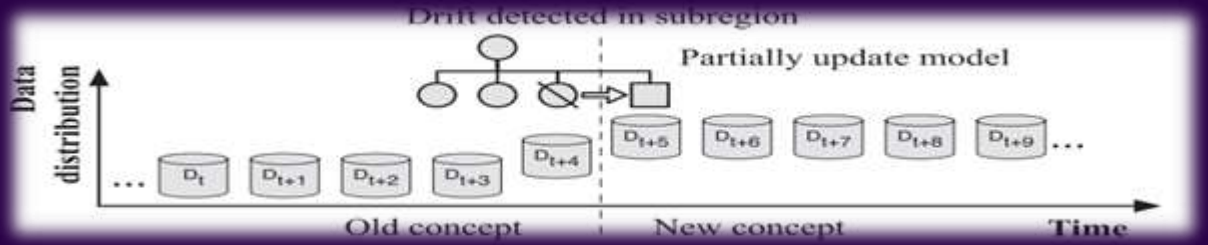
NEW MODEL FOR GLOBAL DRIFT



ENSEMBLE FOR RECURRING DRIFT



ADJUSTING EXISTING MODELS



CONCEPT DRIFT USAGE WITH DIFFERENT AREAS

TEŞEKKÜRLER

Osman İlge Ünaldı
201101006

Arda Erol
201401013