Data Mining Lab: Dynamic Weighted Majority for Incremental Learning

Miena Basta Badres, David Jozefiak

June 10, 2020

Introduction
Concept drifts

- concept drifts occurring in data streams will jeopardize the accuracy and stability
- if the data stream is imbalanced, it will be even more challenging to detect and handle the concept drift
- these two problems have been intensively addressed separately
- they have yet to be well studied when they occur together

DWMIL
Key features

- chunk-based incremental learning method
- deals with data streams with concept drift and class imbalance problem
- creates a base classifier for each chunk
- weighs them by their performance tested on the current chunk
- a classifier trained recently or on a similar concept will receive a high weight

DWMIL
Four major merits

- can keep stable for non-drifted streams and quickly adapt to the new concept
- is totally incremental, no previous data needs to be stored
- keeps a limited number of classifiers to ensure high efficiency
- is simple and needs only one threshold parameter

DWMIL

Method explanation

- on each data chunk $\mathcal{D}(t)$ at timestamp t, a new classifier H is learned
- the new classifier H is merged with $\mathcal{H}(t-1)$ to form the set $\mathcal{H}(t)$
- classifiers are associated with the vector of weights, denoted as $w^{(t)} = [w_1^{(t)}, ..., w_m^{(t)}]^T$
- weights measure the importance of the classifiers in the set
- a weight $w_j^{(t)}$ for classifier $H_j^{(t)}$ is reduced on each timestamp
- the adjusted weight is given by $w_j^{(t)} = (1 - \epsilon_j^{(t)}) \cdot w_j^{(t-1)}$
- finally, new data x is predicted with $sign(\sum_{j=1}^{m} w_j^{(t)} \cdot H_j^{(t)}(x))$

Imbalanced streaming data sets

Real world data sets

- Weather
  - weather information of Bellevue in Nebraska
  - each day can be classified as “rainy” or “not rainy”
- Electricity
  - changes of the electricity price of New South Wales in Australia

Imbalanced streaming data sets

Synthetic data sets

- Moving Gaussian
  - consists of two Gaussian distributed classes
- SEA
  - contains three attributes ranging from 0 to 10
  - only the first two attributes are related to the class
- Hyper Plane
  - contains gradually changing decision hyperplane concepts
- Checkerboard
  - nonlinear XOR classification problem

Imbalanced streaming data sets

Further data sets

- Forest Covertype
  - contains the cover type for 30 x 30 meter forest cells
  - 581,012 instances and 54 attributes
- Poker Hand
  - consists of 1,000,000 instances
  - each instance represents a hand having five poker playing cards
  - each card is described by the attributes suit and rank

The “Weather” data set
Overview

- consists of 18,159 daily readings
  - 5,698 (31 %) are classifed as “rainy”
  - the remaining 12,461 (69 %) are classifed as “not rainy”
- missing values were synthetically generated
- 8 weather features
  - Temperature (Fahrenheit)
  - Dew Point (Fahrenheit)
  - Sea Level Pressure (hPa)
  - Visibility (Miles)
  - Average Wind Speed (Knots)
  - Maximum Sustained Wind Speed (Knots)
  - Maximum Temperature (Fahrenheit)
  - Minimum Temperature (Fahrenheit)

The “Weather” data set
Analysis

- no missing values at all
- imperial units were used, so we converted them into metric units during analysis
- all values are floats, so we can easily calculate **min**, **max**, **mean** and **std** for every value, e. g. for the temperature:
  - min = -24.3 °C
  - max = 33.6 °C
  - mean = 10.6 °C
  - std = 11.7 °C
- two major outliers in the **pressure** column (5503.8 hPa and 5503.1 hPa)

Performance metrics
F1-Score

- measure of a test's accuracy
- considers both the **Precision** and the **Recall** to compute the score

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Performance metrics
Geometric Mean Error

- Geometric Mean is the n-th root of the product of n numbers:

$$\epsilon_{gm} = 1 - \sqrt{TPR \cdot TNR}$$

- True Positive Rate (TPR) or **Recall** / **Sensitivity**:

$$TPR = \frac{TP}{TP + FN}$$

- True Negative Rate (TNR) or **Specificity**:

$$TNR = \frac{TN}{TN + FP}$$

Performance metrics
Area Under Curve (AUC)

- the **ROC curve** is showing the performance of a classification model by plotting TPR and FPR
- the two-dimensional area underneath the ROC curve from (0,0) to (1,1) is called **Area Under Curve (AUC)**
- True Positive Rate (TPR) or **Recall** / **Sensitivity**:

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$