An Investigation of Cross-Project Learning in Online Just-In-Time Software Defect Prediction – Supplementary Material

Sadia Tabassum
sxt901@cs.bham.ac.uk
University of Birmingham, UK

Leandro L. Minku*
L.L.Minku@cs.bham.ac.uk
University of Birmingham, UK

Danyi Feng
danyi@ouchteam.com
Xiliu Tech, China

George Cabral
george.gcabral@ufrpe.br
Federal Rural University of Pernambuco, Brazil

Liyan Song
L.Song.1@cs.bham.ac.uk
University of Birmingham, UK

Algorithm 1 Ensemble approach

S = stream of incoming changes from n projects, b = index identifying the test project, w = waiting period

initialise ensemble model M consisting of n models $\{m_1, m_2, ..., ..., m_n\}$
for each incoming change $x_p^t \in S$ **do** // x_p^t is a change arriving from project p at timestamp t

    if $p = b$ **then**
        $\hat{Y}$-$List \leftarrow$ get prediction $\hat{y}_i$ from each model in M
        $\hat{y}$ = mean of stored prediction results in $\hat{Y}$-$List$
    end if
    store x_p^t in a queue $WFL\text{-}Q_p$ for project p // $WFL\text{-}Q_p$ is the queue of incoming changes of project p waiting to be used for trained

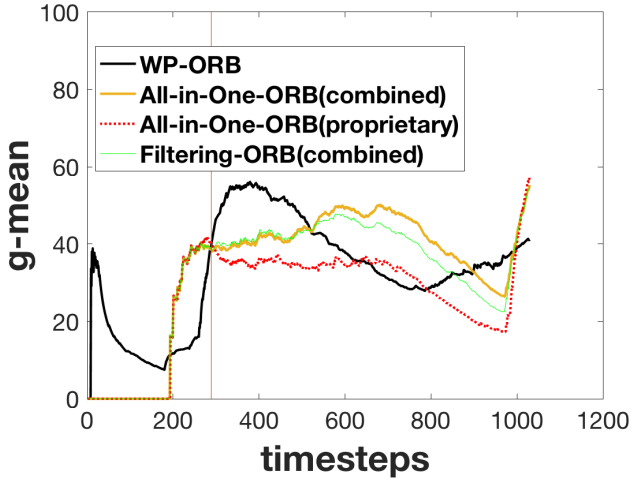
    for each model m_p in M **do**
        for each change q^i in $WFL\text{-}Q_p$ **do**
            if a defect was linked to q^i at a timestamp ≤ t **then**
                create defect-inducing $training\_example$ for q^i
                train(m_p, $training\_example$)
                remove q^i from $WFL\text{-}Q_p$
            else
                create a clean $training\_example$ for q^i
                train(m_p, $training\_example$)
                remove q^i from $WFL\text{-}Q$
                store $training\_example$ in CL_p-H // CL_p-H is a hash of clean training examples for project p
            end if
        end for
    end for

    if a defect was linked to a $training\_example$ in CL_p-H before time t **then**
        Swap the label of $training\_example$ to defect-inducing
        train(m_p, $training\_example$)
        remove h from CL_p-H
    end if
end for

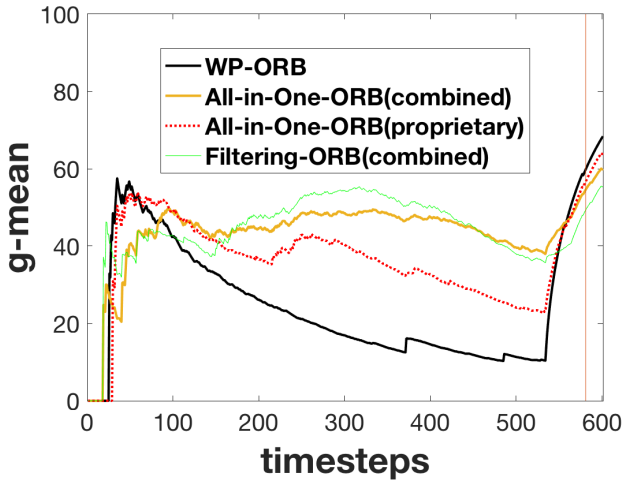
*Corresponding author.

,

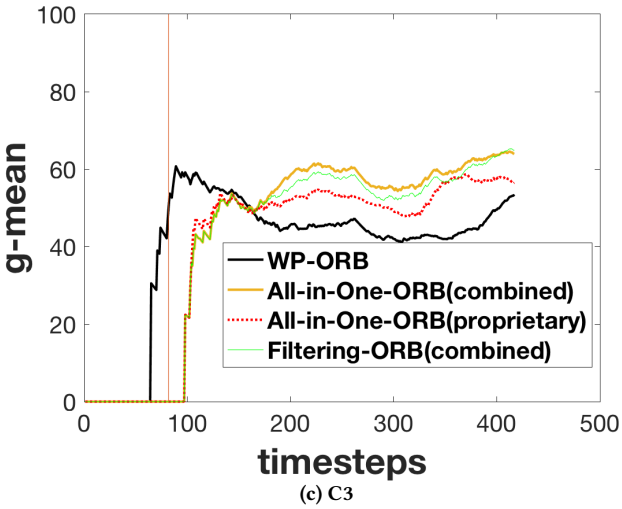
,



(a) C1



(b) C2



(c) C3

Figure 1: G-Mean for proprietary datasets through time using ORB. The vertical red bar indicates the last time step of the initial phase of the project.