IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 3, Ver. VI (May – Jun. 2015), PP 64-69
www.iosrjournals.org

Concept Drift for obtaining Accurate Insight on Process Execution

Vidya N.Chavada[1], Pratik Kumar[2]
[1]M.E Student, Department of CSE, PIET, Vadodara, India
2Asst. Prof.,Department of CSE, PIET, Vadodara, India

Abstract: *Most business processes update with respect time, existing or old process mining strategies used to resolve those processes when they are under the stable position. My attempt is to improve the existing Drift Detection in complex Datasets. Organizations would be interested in discovering the progression of change. There are other applications such as deriving a configurable model for the process change. Online learning algorithms manage concept drifts in process. There are many dissimilar diversity levels in learning machines are needed for keep high generalization. Motivated by this concept, propose a new online learning model named as Diversity for Dealing with Drifts (DDD). By Comparing DDD with other method EDDM conclude that DDD gives higher accuracy. It keeps ensembles with many dissimilar diversity levels and can achieve good accuracy than EDDM. Moreover, it is strong and precise when there are false positive drift detections occur than EDDM. DDD always performed drift handling under many different situations, with very low errors.*

Keywords: *Ensemble learning, Diversity, Concept Drift, DDD.*

I. Introduction

The Although most business processes change with time, existing process mining techniques tend to analyze these processes as if they are in a steady state [3].Processes may be modifying unexpectedly or slowly. There is a situation when the relation between the input data and the output variable, changes with time in unexpected ways which is called concept drift and as in today's dynamic and demanding business, it requires detecting it during online analysis as it although most business processes modify with time [4]. Although detecting concept drifts is important for offline analysis, it is also application oriented and interested to detect changes in near real time (online analysis). Online learning has been showing to be very useful for a growing number of applications in which training data are available continuously in time (streams of data). Examples of such applications are industrial process control, computer security, intelligent user interfaces, market-basket analysis, information filtering. [5]

In this paper used EDDM (Early Drift Detection Method) and compare it with my proposed method DDD.EDDM works on thought that the space between two successive errors increases.so, in this distance is analyze and if it's reduces according to predefined constant then it's measured as concept drift. Diversity with drift detection stored ensembles with altered diversity levels and it is use to attain good precision than EDDM. It is very strong, than any other drift handling techniques in accuracy when there is any false positive change detections occur. DDD used in many change handling approaches in many diverse situation, with very less errors.

In the following section Flowchart of proposed system, Algorithm, experiment result of the system will explain.

II. Flow Chart Of Proposed System

Proposed system for dealing with concept drift consists of the following modules. There are many approaches proposed to handle concept drift. Here I have used EDDM and based upon that proposed the DDD method.

2.1. Dataset Selection and Extraction: in this first of all dataset is selected. In EDDM kc3 dataset and in DDD mc1 dataset is used. In EDDM method I have used parameters like α and β, Where α > β. Value for both parameters determined whether the warning level is triggered and whether the concept drift id detect or not. While in DDD parameters used are pl (low diversity), ph (high diversity), W (multi pliers constant for the weight old low diversity ensemble) default value is (W=1), pd (parameter for drift detection method).

2.2. Calculation of Accuracy: after extracting dataset accuracy of drift detection is calculated for both EDDM and DDD. Both Approaches use some measure related to the accuracy to detect drifts. I have included measures like TP rate, FP rate, precision, and recall, F-measure for false and true class. And also weighted average is calculated. The rule used to obtain the accuracy on time step t is presented in

$$acc(t) = \begin{cases} acc_{ex}(t), & \text{if } t = f, \\ acc(t-1) + \dfrac{acc_{ex}(t) - acc(t-1)}{t-f+1}, & \text{other wise;} \end{cases}$$

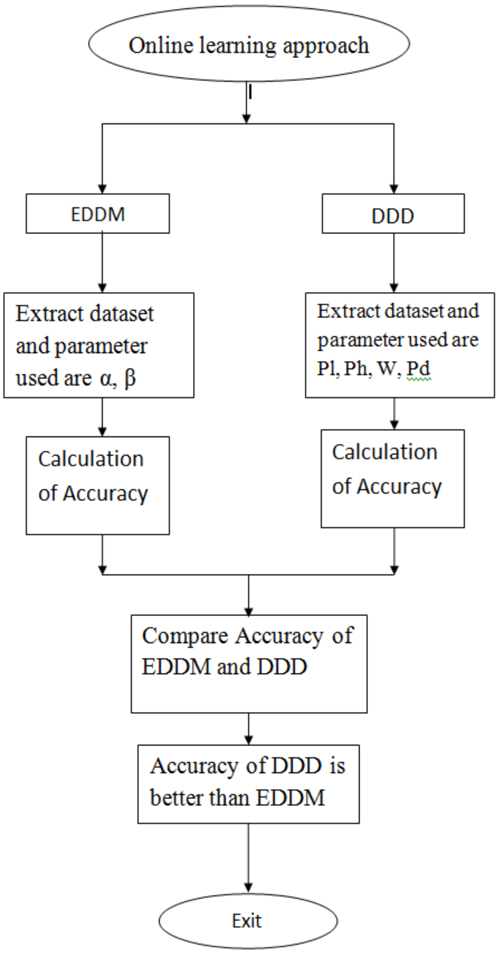

Online learning approach
EDDM
DDD
Extract dataset and parameter used are α, β
Extract dataset and parameter used are Pl, Ph, W, Pd
Calculation of Accuracy
Calculation of Accuracy
Compare Accuracy of EDDM and DDD
Accuracy of DDD is better than EDDM
Exit

Figure. 1: Architecture of DDD

2.3. Compare Accuracy: After that comparison of accuracy of EDDM and DDD is done. And based upon that DDD gives more accuracy than EDDM in both dataset.

III. Description

Diversity with drift detection is in two modes: before the detection of change and after detection of change. We have used a change detection method, because it permits instant behaviour of changes when they are found. So, if the parameters of the change detection method are coping to detect change the first possible and the approach is considered to be strong to detect the false alarms, we may attain quick adaptation to novel concepts.

Online analysis is viewing to be extremely helpful for an increasing number of applications in that training information are obtainable constantly with respect to time. Example of such applications is process control in organization, security of computer, market-basket analysis, filtering of information.

In online analysis algorithms it process each training example one time “when coming” with no need of storage space. In this work we have done a method for detecting concept drift, even in the case of slow gradual change. It is based on the estimated distribution of the distance between classification errors.

IV. Algorithm

Inputs:

- multiplier constant W for the weight of the old low diversity ensemble;
- online ensemble learning algorithm EnsembleLearning;
- parameters for ensemble learning with low diversity pl and high diversity ph;
- .drift detection method DetectDrift;
- parameters for drift detection method pd;
- data stream D;

```
mode < - before drift
hnl <- new ensemble /*new low          diversity*/
hnh < - new ensemble /*new high diversity*/
hol < - hoh < - null /*old low and high diversity*/
accol < - accoh < - accnl < - accnh < - 0 /*accuracies*/
stdol < - stdoh < - stdnl < - stdnh <- 0 /*standard deviations*/
while true do
  d < - next example from D
  if mode == before drift then
    prediction hnl(d)
  else
    sumacc < - accnl + accol* W +accoh
    wnl =accnl/sumacc
    wol = accol *W/sumacc
    woh = accoh/sumacc
    prediction<- WeightedMajority(hnl(d), hol(d),hoh(d), wnl, wol, woh)
    Update(accnl, stdnl, hnl, d)
    Update(accol, stdol, hol, d)
    Update(accoh, stdoh, hoh, d)
  end if
  drift<- DetectDrift(hnl, d, pd)
  if drift == true then
    if mode == before drift OR(mode == after drift AND accnl > accoh)  then
      hol <-hnl
    else
      hol<- hoh
    end if
    hoh<- hnh
    hnl <- new ensemble
    hnh <- new ensemble
    accol <- accoh <- accnl <- accnh <- 0
    stdol <- stdoh <- stdnl <- stdnh <- 0
    mode <- after drift
  end if
  if mode == after drift then
    if accnl > accoh AND accnl > accol then
      mode<- before drift
    else
      if accoh - stdoh > accnl +stdnl AND accoh - stdoh > accol + stdol then
        hnl<- hoh
        accnl<- accoh
        mode<- before drift
      end if
  EnsembleLearning(hnl; d; pl)
  EnsembleLearning(hnh; d; ph)
  if mode == after drift then
    EnsembleLearning(hol; d; pl)
    EnsembleLearning(hoh; d; pl)
  end if
  if mode == before drift then
```

```
51: Output hnl, prediction
52: else
53: Output hnl; hol; hoh; wnl; wol; woh, prediction
```

Before a detection of drift, the learning system is combined with two ensembles: lower diversity (hnl) ensemble and higher diversity (hnh) ensemble. Both hnl and hnh are trained with input samples (lines 44 and 45), here we are considering only low hnl ensemble for predictions of the system (line 10). Hnh is not use for prediction of the system. Because it's accuracy is low on the new concept. A change detection method for monitoring the hnl is used (line 21). After a detection of drift, hnl and hnh is generated (lines 29 and 30). Before the drift is detected Low and high diversity ensembles are reserved and treated diversity Ensembles as old low and old high (lines 24 and 28). Learning is done by Both the old and the new ensembles (lines 44-48) and then calculations of the system are determined by the weighted majority vote of the output of 1) hol, 2) hnl, and 3) hoh (lines 12-16).

The weights are proportional to the accuracy since the last detection of drift until the preceding time step (lines 13-15). The weight of the hol ensemble is multiplied by a constant W (line 15), that controlling between strength to false alarms detection and precision in the existence of concept drifts, and then normalization of weights are performed. If two successive detections of drift occur, the hol ensemble after the detection of the second drift can be either the equal as the hoh learning with low diversity after the detection of the first drift or the ensemble equivalent to the hnl after the first detection of the drift, depends on which of them more precise (lines 24 and 26). All the four ensembles are stored until either the condition 36 or the condition 39 is fulfilled.

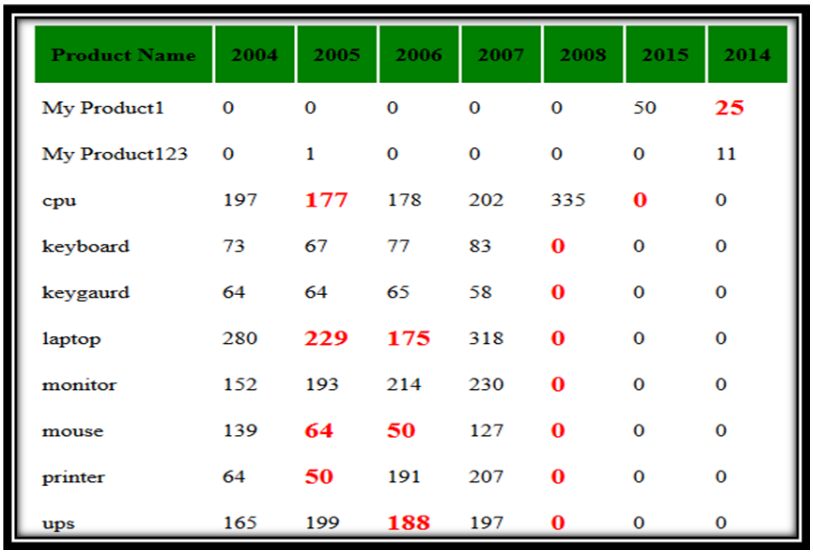
When considering the accuracies whether the hoh ensemble is better than the others are reduced/summed from their standard deviations.

While returning back to the mode earlier to drift, either the hoh or the hnl ensemble becomes the low diversity ensemble used in the mode earlier to detection of drift, depending on which of them is more precise. (lines 36-43).

Diversity in Detection Drift is considered to use data well-read from the old theory in order to help the learning of the original theory by training a hoh ensemble with low diversity on the new theory.

V. Experimental Result

In existing system for evaluating the concept drift in data set, manually data set of different electronic hardware items is taken.



Product Name	2004	2005	2006	2007	2008	2015	2014
My Product1	0	0	0	0	0	50	25
My Product123	0	1	0	0	0	0	11
cpu	197	177	178	202	335	0	0
keyboard	73	67	77	83	0	0	0
keygaurd	64	64	65	58	0	0	0
laptop	280	229	175	318	0	0	0
monitor	152	193	214	230	0	0	0
mouse	139	64	50	127	0	0	0
printer	64	50	191	207	0	0	0
ups	165	199	188	197	0	0	0

Figure. 2: Concept Drift in Product in each year

So, as shown in Fig..2 it's offline analysis of concept drift in artificial dataset. In this characterization of the concept drift is not done.

In proposed system we have taken PROMISE data set in order to encourage repeatable, verifiable, refutable, and/or improvable predictive models of Diversity for Dealing with Drifts. Two real time data ser mc1.arff and kc3.arff is taken.

By performing evaluation of both dataset in to the system accuracy achieved by DDD and EDDM for different dataset is shown in following table:

Table 1: Comparison Table EDDM and DDD

Dataset Name	EDDM	DDD
kc3.arff	84.7826	86.9565
mc1.arff	99.472	99.5248
kc3.arff,mc1.arff	84.7826	99.5248
mc1.arff, kc3.arff	99.472	86.9565

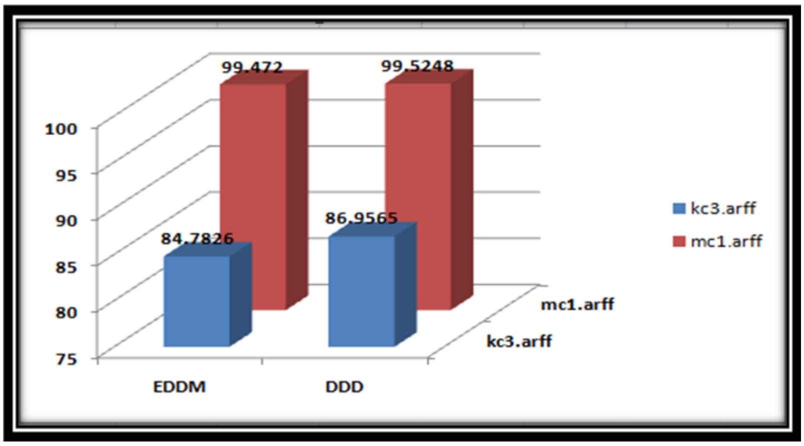


Figure .3: Comparison of EDDM and DDD for one dataset based on accuracy

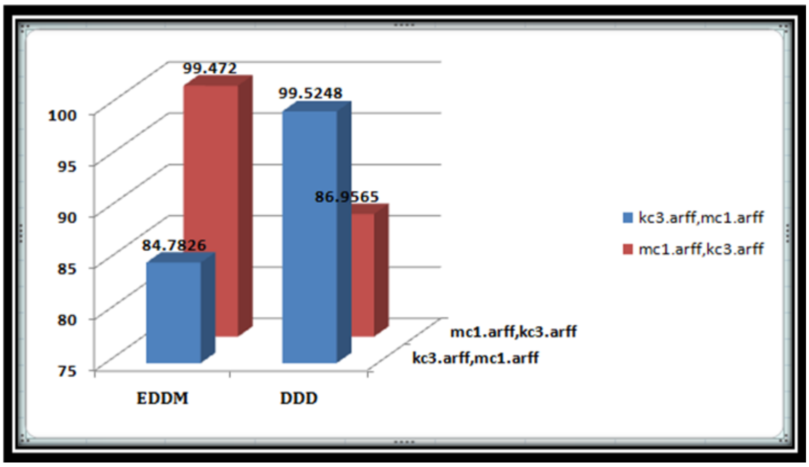


Figure .4: comparison of EDDM and DDD for two different dataset based on accuracy.

VI. Comparison Analysis Of Proposed And Existing System

After Successfully Evaluate the result of Proposed Approach with compare to Existing Approach, we will see a comparative study of both system.

Table 2: Comparison Table of Proposed and Existing system

Proposed System	Existing System
• **Its online analysis because I have used real-time dataset and then compare dataset for EDDM and DDD method.**	• Its offline analysis because I have use artificial data.
• **Real time dataset is used.** • **Data set available in order to encourage repeatable, verifiable, and/or improvable predictive models of Diversity for Dealing with Drifts.** • **More than one dataset is used.**	• Artificial dataset is used. • And manually calculated the drift in product and year wise. • One dataset is used
• **It will automatically find false positive and false negative drift. Precision is also measure.**	• Put manually drift and then check
• **Accuracy of EDDM and DDD is calculated. And DDD gives more accuracy than EDDM.**	• Accuracy for drift is not calculated.
• **Characterization is done in graph format for checking drift.** • **Diversity is also considered.**	• Characterization is not done.

VII. Conclusion

In proposed work taking more than one real-time dataset and analyze and characterize the drift by comparing EDDM and DDD method. Online Analysis will use Diversity with Drift Detection approach for finding drift in data set. After Successfully Evaluate the result of Existing Approach, the further proposed work will find the drift in data set with diversity also, using DDD approach. In experimental result we take two different dataset and compare the EDDM and DDD for different dataset. So, from the result we can conclude that DDD gives better result than EDDM in finding the drift. And using that approach accuracy of detecting drift will increase compare to existing method. Diversity with drift detection has also good strength not in the favor of false alarms. As and when they occur, its precision is good due to the usage of old ensembles. So, Diversity with Drift Detection is precise both in the existence and in the nonexistence of drifts compared to EDDM. Therefore, this analysis should only be observed as the starting point for a new subfield in the concept drift domain.

References

[1]. Manoj Kumar M V,Likewin Thomas, Annapa B, "Phenomenon Of Concept Drift From Process Mining Insight"Department Of CSE,NITK,978-1-4799-2572-8/14/$31.00 2014 IEEE.

[2]. Stb Prof. Dr. Nick Gehrke Nordakademie "Process Mining" Chair For Information Systems Köllner Chaussee 11,Michael Werner, Dipl.-Wirt.-Inf. University Of Hamburg Chair For Information Systems , Max-Brauer-Allee 60 D-22765 Hamburg.

[3]. R. P. Jagadeesh Chandra Bose, Wil M. P. Van Der Aalst, Indr˙E Žliobait˙E, And Mykola Pechenizkiy, "Dealing With Concept Drifts In Process Mining"**,** Ieee Transactions On Neural Networks And Learning Systems, Vol. 25, No. 1, January 2014.

[4]. Online Techniques For Dealing With Concept Drift In Process Mining, Josep Carmona And Ricard Gavalda, Universitat Polit_Ecnica De Catalunya Barcelona, Spain 2013.

[5]. J. Carmona And R. Gavaldà, "Online Techniques For Dealing With Concept Drift In Process Mining," In Proc. Int. Conf. IDA, 2012, Pp. 90–102.