ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

DELA: A Dynamic Online Ensemble Learning Algorithm

Abdelhamid Bouchachia and Emili Balaguer-Ballester

Faculty of Science and Technology, Bournemouth University, UK

Abstract. The present paper investigates the problem of prediction in the context of dynamically changing environment, where data arrive over time. A Dynamic online Ensemble Learning Algorithm (DELA) is introduced. The adaptivity concerns three levels: structural adaptivity, combination adaptivity and model adaptivity. In particular, the structure of the ensemble is sought to evolve in order to be able to deal with the problem of data drift. The proposed online ensemble is evaluated on the stagger data set to show its predictive power in presence of data drift.

1 Introduction

Online learning has received a great attention from the machine learning community. Its origin goes back to the late 1980s and early 1990s with the advent of the paradigm of "prediction with expert advice". The early work appeared in a number of seminal papers [5]. Online learning consists of learning a sequentially presented set of training data upon arrival, without re-examining data that have been processed so far. Online learning is practical for applications where the data set is large and cannot be processed at once due to memory constraints. Practically an online learner OL receives a new data point x_t along with the current hypothesis H_{t-1}, checks if the data point is covered by the current hypothesis and updates the hypothesis accordingly. Ideally the online learner performs as well as its corresponding offline version, where it can see the data for many epochs. If this happens we say that the algorithm is lossless. The protocol of online learning can be summarized as follows:(1) the learner receives an observation x_t, (2) the learner makes a decision $\hat{y}$, (3) the learner receives the ground truth y, (4) the learner incurs a loss $\ell(\hat{y}; y)$ and updates its hypothesis w.

One of the earliest successful implementation of this protocol is the mistake-driven algorithm known as the Weighted Majority Algorithm (WMA) [5]. WMA combines online learning with ensemble learning in a way that data are presented sequentially over time and where each expert (predictor) makes a decision. After that, the weighted majority voting technique is applied in order to produce the final output. The weights of the experts change according to their individual performance. In particular, the weights of the learners, reflecting the number of mistakes that have been made so far, are updated when the ensemble misclassifies the current input following the original scheme of the Winnow algorithm. The experts that produce the right predictions (correct decision) are promoted (by increasing their weights) and those that fail to predict the correct actual output are demoted (i.e. penalized) by decaying their weight. Due to this weighing mechanism, the main advantage of WMA is that it can perform as good as the best of the individual experts at any time. Moreover, it has been shown that it can cope with the problem of data drift [4] but in a context where the expert ensemble is dynamic.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The present paper introduces a weighted majority voting algorithm, called "Dynamic online Ensemble Learning Algorithm" (DELA) for coping with dynamically changing environment such as concept drift.

The paper is organized as follows. In Sec. 2 a short overview of the state-of-the-art of online learning, ensemble learning and concept drift is presented. Section 3 describes the algorithm DELA with ample details. Section. 4 provides the initial simulation results. Finally Sec. 5 concludes the paper.

2 Short overview of the literature

One of the most successful online algorithms that involves a pool of experts is the weighted majority algorithm (WMA) proposed by Littlestone and Warmuth [5] which is inspired from the halving algorithm. In WMA the weights of the learners in the pool are initialized to 1. Relying on the multiplicative learning rule (like in the Winnow), the weights are multiplied by a parameter $0 < \beta < 1$ such that the contribution (weight) of the learners that make a large number of mistakes gets decayed in contrast to those learners that make fewer mistakes.

Learning with expert advice has been modeled as a game theory problem, where the aggregation of the experts is investigated. Weight of the experts are seen as a function of the learning loss observed over the past trails. The goal is to have the number of mistakes of the algorithm as high as those made by the best expert.

Quite interestingly the problem of drift handling has been discussed in [3] by looking at segments of the data and finding the experts that perform better on such segments. However, the algorithm proposed does not produce exactly the optimal assignment expert-segment, but it produces near-optimal assignment.

In [4] a modified version of WMA, called dynamic weighted majority algorithm (DWM), is proposed. The algorithm enables adding and deleting new experts on the fly following certain conditions. The underlying motivation of DWM is that such operations help in dealing with learning of drifting concepts. A similar algorithm, equipped with the operations addition and deletion of experts, is proposed in [1]. Additional operations are also considered: promotion of the expert whenever the outcome is correctly predicted and demotion whenever an incorrect prediction is made.

In connection to data drift and online learning, the application of ensemble learning has been the subject of few investigations over the recent past years. For instance, in [2] a batch-based ensemble of classifiers is proposed to deal with concept drift. In [6] diversity of learning ensemble is investigated in presence of different types of drifts.

In this paper, we present DELA which is characterized by dynamic combination of online experts and by dynamic and continuous structural update of the ensemble over time as soon as the accuracy of the ensemble starts to deteriorate. In a nutshell, the DELA approach is distinct compared to other studies [4] in the following aspects: (1) Base experts are distinct, (2) Base experts are adaptive and incremental by their nature, (3) Base experts handle novelty detection by their nature, (4) Ensemble method implements a dynamic combination of experts and new classifiers may be added while existing ones may be removed, (5) Ensemble method proposed deals with data drift, (6) Adaptation takes place at three levels: expert, combination, structure of the ensemble.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

3 Description of DELA

DELA is based on the Weighted Majority Algorithm (WMA) and the Winnow algorithm which have been explained in the previous section. The ensemble consists of a set of experts that operate online and point-wise in contrast to most of the state-of-the-art techniques which operate on a window basis.

Each ith expert, ζ_i, $i = 1 \cdots M$ makes a prediction $p_i(t) \in \mathcal{C}$ which is a set of labels (i.e., prediction space). The ensemble derives a prediction $p(t) \in \mathcal{P}$ using $p_i(t), i = 1 \cdots M$ using the weighted majority voting since each expert is associated a corresponding weight. The correct label π_t is then presented. The experts are promoted, i.e., their weights are increased, if their predictions turn to be correct, otherwise they are penalized, i.e., their weights are decreased.

Being dynamic, DELA is equipped with two operations:

1. Addition of experts: an expert is added when the following conditions (*Add_Conditions*) are satisfied:
   - the ensemble fails to predict the correct label
   - the ensemble has made so far as many mistakes as half of a window of length Q
   - the number of experts currently in the pool is less than a certain threshold M
2. Removal of experts: an expert is deleted from the ensemble pool when the following conditions (*Delete_Conditions*) are met:
   - the learning ensemble fails to predict the correct label
   - the expert to be removed has made a mistake on each trail over the last window (of length Q)
   - the number of experts in the current pool is at least 2

The creation of a new expert obeys a rule stipulating that the new expert should keep (or increase) the diversity of the ensemble. To achieve this, we rely on the historical performance of the experts to measure the discrepancy between the experts. The expert with high discrepancy is selected. We consider a maximum number of experts M from a type set Θ: Kernel Passive-Aggressive multiclass, Kernel Perceptron/Random Budget Perceptron multiclass, and Kernel Projectron++ multiclass [7].

The discrepancy between two experts, ζ_1 and ζ_2, is given as follows:

$$dis(\zeta_1, \zeta_2) = \sum_{\tau=t_0}^{t} [p_1(\tau) \neq p_2(\tau)] \tag{1}$$

where the current time window is $[\tau_0, t = Q + \tau_0]$. The diversity of an expert compared to the rest of the experts is given by:

$$div(\zeta_i) = \sum_{j=1; j \neq i}^{M} dis(\zeta_i, \zeta_j) \tag{2}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The diversity of an expert is the cumulative count of prediction discrepancy with respect to the other experts. Therefore the type θ of the expert to be added will be:

$$\theta = \underset{i=1:M}{\arg\max}\, div(\zeta_i) \tag{3}$$

Finally, DELA consists of the steps shown in Alg. 1. For the sake of understandability, the steps are kept as short as possible especially in lines 14 and 21, described in the text.

4 Experimental Simulations

To check the performance of DELA, we have used two data sets presenting two types of concept drift: sudden and gradual. The motivation behind using this data set is to show how DELA can react in critical situations, that is in presence of drift. The first data set is the standard stagger data and simulates sudden drift. To understand the behavior of DELA, each time an example from a concept is presented. Then, a sample of 100 randomly generated examples from the same concept as the current training example is used to evaluate the performance of the algorithm. That is, after each training step, the accuracy of the algorithm is computed. The experiments below are repeated 30 runs and the accuracy results are averaged. The second dataset is an oscillatory multivariate time series, where the periodic behaviour subtly changes over time. Data consists of hourly ozone and nitric oxides ground concentrations, as well as temperature and relative humidity recordings averaged over 60 min intervals during 8 weeks.

The experts used by DELA in this simulation are provided in [7]: Kernel Passive-Aggressive multiclass, Kernel Perceptron/Random Budget Perceptron multiclass, and Kernel Projectron++ multiclass. It is important to note the diversity and the nature of these algorithms which are all incremental and operate online. The DELA's parmeters: promotion (α), demotion β) and window length (Q) are set to 1.3, 0.6 and 15 respectively. Figure 1 shows the evolution of the accuracy over time on the Stagger dataset. Clearly for sudden drift DELA adapts easily, every time a new concept is introduced the classifier does not take long to recognize it. This behavior is similar to that of the DWM classfier as shown in Fig.2.

When tested on the Ozone dataset which presents gradual drift, DELA and DWM exhibt a high accuracy but DELA outperforms DWM as illustrated in Fig. 3. The number of experts tends however to be very dynamic in DWM, while with DELA the number increases over time.

5 Conclusion

The present paper is concerned with the development of a new online ensemble learning algorithm, called DELA, dedicated to dynamically changing environments. Inspired from the well known weighted majority algorithm, it is equipped with the addition and deletion of experts operations, which can be which can be executed dynamically and just-on-demand. The first simulations show that the algorithm is able to quickly adapt to drift. It is also worth mentioning that it is easy to show that DELA is bounded in terms of error. Due to space, this analysis was left out. Further investigations still need to be done on the effect of diversity, cyclic drift and large data sets.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Algorithm 1 DEL algorithm

Given a pool of experts Ψ of different types that can be integrated in the ensemble, let S be the initial number of experts to start with, $S \in \{1, \cdots, |\Psi|\}$
Set the parameters: Q length of the observation window (used to track the accuracy of the experts); M the maximum number of experts in the ensemble
Initialize weights $w_1, ..., w_S$ of the $S < M$ experts, the promotion parameter α ($\alpha > 1$) and the demotion parameter β ($0 < \beta < 1$), the weights w_i s.t. $\sum_i^S w_i = 1$
At time t:
Present the current input $x(t)$ to each expert ζ_i
Get predictions from the experts $(p_1(t), \cdots p_R(t))$ where $R \leq M$ is the current number of experts in the ensemble
Compute the weighted majority vote (the decision of the ensemble)

$$\tilde{y}(t) = \underset{j,\ j=1\cdots\mathcal{C}}{\arg\max} \left(\sum_{i=1}^{R} w_i [c_j = p_i(t)] \right), \text{where } \mathcal{C} \text{ is the set of labels}$$

Compute the diversity of the experts using Eq. 2
if $\tilde{y}(t) \neq y(t)$ **then**
  for all Experts $\zeta_i,\ i = 1 \cdots R$ **do**
    if $p_i(t) = y(t)$ **then**
      $w_i(t+1) = w_i(t) * \alpha$
    else
      if Delete_Conditions as described in the text are met **then**
        Delete the expert ζ_i
      else
        $w_i(t+1) = w_i(t) * \beta$
      end if
    end if
  end for
  if Add_Conditions as described in the text are met **then**
    Add an expert k of type θ (Eq. 3)
    Set its weight w_k to 1
  end if
end if
Normalize the weights:

$$w_i(t+1) = \frac{w_i(t+1)}{\sum_{j=1}^{n} w_j(t+1)}, i = 1 \cdots R$$

Train each expert $\zeta_i,\ i = 1 \cdots R$ on the the input $< x(t), y(t) >$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

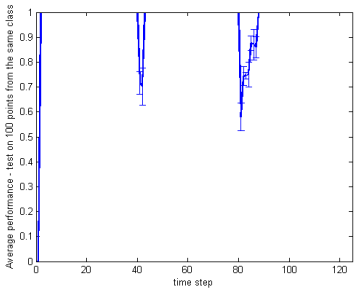

Fig. 1: Current accuracy of DELA

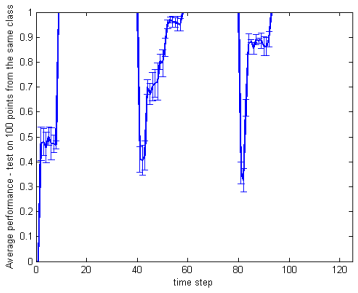

Fig. 2: Current accuracy of DWM

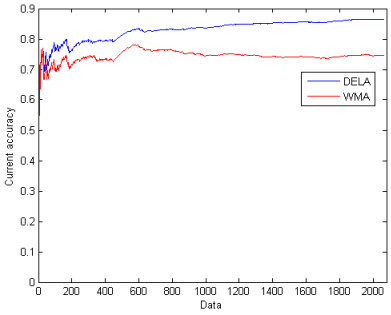

Fig. 3: Current accuracy: DELA vs. DWM

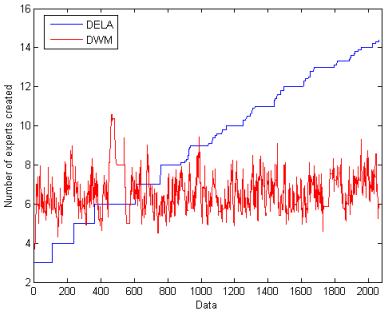

Fig. 4: Number of experts generated

References

[1] A. Bouchachia. Incremental learning with multi-level adaptation. *Neurocomputing*, 74(11):1785–1799, 2011.

[2] R. Elwell and R. Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 2011.

[3] M. Herbster and M. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, 1998.

[4] J. Kolter and M. Maloof. Dynamic weighted majority: An ensemble method for drifting concepts. *Journal of Machine Learning Research*, 8:2755–2790, 2007.

[5] N. Littlestone and M. Warmuth. The weighted majority algorithm. *Inf. Comput.*, 108(2):212–261, 1994.

[6] White A. Yao X. Minku, L. The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Transactions on Knowledge and Data Engineering*, 22(5):730 –742, may 2010.

[7] F. Orabona. Dogma: a matlab toolbox for online learning. Software available at http://dogma.sourceforge.net, 2009.