The Impact of Diversity on On-line Ensemble Learning in the Presence of Concept Drift

Leandro L. Minku, *Student Member, IEEE,* Allan P. White, and Xin Yao, *Fellow, IEEE*

Abstract—On-line learning algorithms often have to operate in the presence of concept drift (i.e., the concepts to be learnt can change with time). This paper presents a new categorization for concept drift, separating drifts according to different criteria into mutually exclusive and non-heterogeneous categories. Moreover, although ensembles of learning machines have been used to learn in the presence of concept drift, there has been no deep study of why they can be helpful for that and which of their features can contribute or not for that. As diversity is one of these features, we present a diversity analysis in the presence of different types of drift. We show that, before the drift, ensembles with less diversity obtain lower test errors. On the other hand, it is a good strategy to maintain highly diverse ensembles to obtain lower test errors shortly after the drift independent on the type of drift, even though high diversity is more important for more severe drifts. Longer after the drift, high diversity becomes less important. Diversity by itself can help to reduce the initial increase in error caused by a drift, but does not provide a faster recovery from drifts in long term.

Index Terms—Concept drift, on-line learning, neural network ensembles, diversity.

1 Introduction

On-line learning has been showing to be very useful for a growing number of applications in which training data is available continuously in time (streams of data) and/or there are time and space constraints. Examples of such applications are industrial process control, computer security, intelligent user interfaces, market-basket analysis, information filtering, prediction of conditional branch outcomes in microprocessors and RoboCup.

On-line learning algorithms process each training example once "on arrival", without the need for storage or reprocessing, and maintain a current hypothesis that reflects all the training examples so far [1]. In this way, the learning algorithms take as input a single training example as well as a hypothesis and output an updated hypothesis [2]. We consider on-line learning as a particular case of incremental learning. The later term refers to learning machines that are also used to model continuous processes, but are allowed to process incoming data in chunks, instead of having to process each training example separately [3].

Ensembles of classifiers have been successfully used to improve the accuracy of single classifiers in on-line and incremental learning [1]–[5]. However, on-line environments are often non-stationary and the variables to be predicted by the learning machine may change with time (concept drift). For example, in an information filtering system, the users may change their subjects of interest with time. So, learning machines used to model these environments should be able to adapt quickly and accurately to possible changes.

Several approaches have been proposed to handle concept drift over the last few years. For instance, we can cite ensemble approaches which create a new classifier to each new chunk of data and weight classifiers according to their accuracy on recent data [6]–[9], possibly using a boosting-like mechanism for the learning [7]–[9]. Another example is Gao et al.'s work [10], which proposes the use of unweighted ensembles, as new data may belong to a concept different from the most recent training data. Street and Kim [11] also report that no consistent improvement on the accuracy was obtained when using ensemble member weights in their approach.

Differently from the approaches that maintain classifiers which learnt old concepts in the ensemble, some approaches discard classifiers when they become inaccurate or when a concept drift is detected [12]–[15].

Although approaches which maintain old classifiers in an ensemble might be claimed to handle recurrent drifts (return to previous concepts), some approaches try to deal with recurrent drifts more explicitly. Ramamurthy and Bhatnagar's approach [16], for example, maintains a global set of classifiers with weights based on their accuracy on the last training chunk of data and uses for prediction only classifiers with error below a certain value. Forman [17], on the other hand, uses old classifiers to generate extra features for the training examples of the current chunk of data. These features are the predictions that the old classifiers would have made for those training examples.

Many incremental learning approaches tend to give less attention to the stability of the classifiers, giving more emphasis to the plasticity when they allow only a new ensemble member to learn a new chunk of data. While this could be desirable when drifts are very fre-

L. Minku and X. Yao are with the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA), School of Computer Science, The University of Birmingham, e-mail: {F.L.Minku,X.Yao}@cs.bham.ac.uk. A. White is head of the Statistical Advisory Service and also an associate member of the School of Mathematics and Statistics at the University of Birmingham, Edgbaston, Birmingham B15 2TT, UK, e-mail: A.P.White@bham.ac.uk.

quent, it is not a good strategy when drifts are not so frequent. On-line learning approaches tend to give more importance to stability when the concept is stable, allowing the whole system to learn all the new incoming data when it is believed that there is no concept drift [13]–[15], [18], [19]. However, some incremental approaches are more careful with this issue as well. For instance, Scholz and Klinkergerg [8] give the possibility for old classifiers to learn new data and Fan [20], [21] allows the use of old training examples to train new classifiers, in case they can help to increase the accuracy of the system.

Some approaches also try to handle skewed distributions, which are frequently present in data stream learning [22].

Although ensembles have been used to handle concept drift, the literature does not contain any deep study of why they can be helpful for that and which of their features can contribute or not to deal with concept drift. The primary goal of the study presented in this paper is to gain a deeper understanding of when, why and how on-line ensemble learning can help to deal with drifts.

In off-line mode, diversity among base learners is an issue that has been receiving lots of attention in the ensemble learning literature. Many authors believe that the success of ensemble algorithms depends on both the accuracy and the diversity among the base learners [23], [24]. However, no study of the role of diversity in the presence of concept drift has ever been done.

As diversity could be a feature that contributes to deal with concept drift, we conduct a study of diversity when working with concept drift classification problems. The study aims at determining the differences between the role of diversity before and after a drift, whether diversity by itself can provide any advantage to handle concept drift and the effect of diversity on different types of drift.

We show that, before the drift, ensembles with less diversity obtain lower test errors. On the other hand, it is a good strategy to maintain highly diverse ensembles to obtain lower test errors shortly after the drift independent on the type of drift, even though high diversity is more important for more severe drifts. After a large number of time steps have passed since the beginning of the drift, high diversity becomes less important. Diversity by itself is helpful to reduce the initial drop in accuracy that happens right after a drift, but not to provide convergence to the new concept.

Besides the diversity study, we also propose a new concept drift categorization, as the categorization existing in the literature is composed of very heterogeneous and somewhat vague categories. The new categorization uses several criteria to divide drifts into mutually exclusive and well-defined categories, allowing more systematic and detailed studies of drifts. We also suggest measures to characterize the drifts according to different criteria and make available a data sets generator that allows the construction of data sets containing different types of drift.

This paper is further organized as follows. Section 2 defines concept drift. Section 3 presents the proposed categorization. Section 4 contains a discussion about previous work on diversity and the importance of studying diversity in the presence of concept drift. Section 5 explains the study of diversity in concept drift. Section 6 presents the conclusions and future works.

2 Concept Drift Definition

We are considering that the term *concept* refers to the whole distribution of the problem in a certain point in time [25], being characterized by the joint distribution $p(\mathbf{x}, w)$, where $\mathbf{x}$ represents the input attributes and w represents the classes. So, a *concept drift* represents a change in the distribution of the problem [14], possibly being a feature change (change only in the unconditional probability distribution function (pdf)), a conditional change (change only in the posterior probabilities) or a dual change (both in the unconditional pdf and in the posterior probabilities) [10].

3 Proposed Drifts Categorization

Different types of environmental changes (and concept drifts) can be identified in the literature. For example, in [14], several artificial data sets are considered to contain abrupt or gradual drifts. The well known SEA Concepts [11] and STAGGER Concepts [26] data sets are considered to have a gradual drift and abrupt drifts, respectively. According to [25], gradual changes have also been called gradual drifts, evolutionary changes or simply concept drifts. Abrupt changes have also been called substitutions, concept substitutions, revolutionary changes, abrupt drifts or concept shifts. Moreover, there are also changes called recurring trends or recurring contexts and population drifts.

Besides the existence of different denominations to the same types of drift, the literature considers only 2 criteria (speed and recurrence), forming very heterogeneous categories. For example, consider the concept drifts shown in figure 1. In this figure, the concept is represented by grey color when the target class is 1. In the literature, both the drifts 1(a) and 1(b) are considered gradual. The drift 1(a) is considered gradual because the new concept starts gradually to take over, while the drift 1(b) is considered gradual because the old concept gradually moves in the direction of the new concept, creating intermediate concepts. However, these drifts are very different from each other.

Moreover, if the concepts I1-3 in figure 1(b) are considered intermediate concepts, the drift is considered gradual. However, as there is no definition of intermediate concept in the literature, the category "gradual drift" is vague and we could consider that there are 4 abrupt drifts, instead of 1 gradual drift.

As it can be seen, the literature needs a less heterogeneous categorization, dividing drifts into different

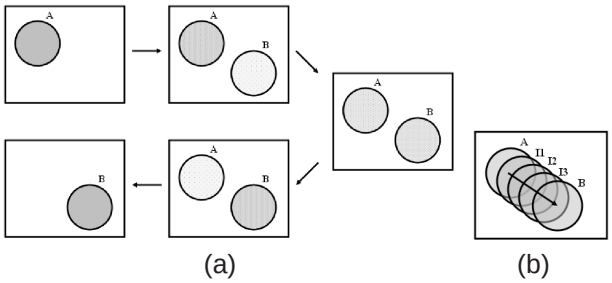

Fig. 1. Heterogeneous categories - gradual drifts.

types, according to different criteria, and creating mutually exclusive and non-vague categories considering a particular criterion. So, we propose a new categorization for concept drift, seeking inspiration from the dynamic optimisation problems area [27]–[29]. We also suggest measures to characterize drifts according to the criteria.

When analysing a drift in isolation, we can observe that different drifts can cause different amount of changes and take more or less time to be completed. For example, in an information filtering system, a reader may slowly change her/his subjects of interest from a particular subject to another or may quickly get interested in an additional subject. The new subject could be very different or could have some similarities to previous subjects, causing more or less changes to the concept. So, in order to categorize a particular drift, we propose to use the criteria severity (amount of changes that a new concept causes) and speed (the inverse of the time taken for a new concept to completely replace the old one), as shown in table 1.

Although there are applications in which drifts may be completely random and without any pattern, there are also applications in which drifts have certain tendencies. Taking again the information filtering system example, a reader may not change her/his general preferences all the time without any tendency, so drifts may be less frequent. Besides, a reader could start reading new different subjects and afterwards loose interest for these new subjects, returning to previous concepts [17]. In weather forecasting, electricity consumption prediction or market basket analysis, there may be drifts which happen depending on the time of the year, having periodic, recurrent and predictable behaviours. The same may happen for prediction of bacterium resistance to antibiotics [30], in which some interesting findings were done about bacterium resistance, such as seasonal context recurring with winter and spring models. So, in order to categorize a sequence of drifts, we propose to use the criteria predictability (whether the drifts are completely random or follow a pattern), frequency (how frequently drifts occur and whether they have a periodic behaviour) and recurrence (possibility to return to old concepts), as shown in table 1.

This section is further organized as follows: section 3.1 explains the criteria to categorize drifts in isolation, section 3.2 explains the criteria to categorize drift sequences

TABLE 1
Concept Drift Categorization

Criteria			Categories	
Drift in Isolation	Severity	Class	Severe	
			Intersected	
		Feature	Severe	
			Intersected	
	Speed		Abrupt	
			Gradual	Probabilistic
				Continuous
Drift Sequences	Predictability		Predictable	
			Non Predictable	
	Frequency		Periodic	
			Non-Periodic	
	Recurrence		Recurrent	Cyclic
				Unordered
			Not Recurrent	

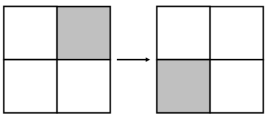

Fig. 2. Example of an intersected drift.

and section 3.3 gives some additional explanation about how to use the proposed categorization, including comments on the use of the term intermediate concept. The proposed categories are summarized in table 1.

3.1 Criteria to Categorize Drifts in Isolation

The criterion severity has not been used in the literature yet. It is further divided into class and feature severity. According to class severity, drifts can be divided into severe and intersected. Drifts are severe when no example maintains its target class in the new concept, i.e., 100% of the input space changes the target class when the drift occurs. If part of the input space has the same target class in the old and new concepts, the drift is intersected. We suggest 2 different measures to characterize drifts according to class severity:

1) Percentage of the input space which has its target class changed after the drift is complete. For example, in the drift represented by figure 2, $2/4 = 50\%$ of the input space has its target class changed.
2) Maximum percentage of the input space associated to a particular class in the old concept that has its target class changed in the new concept. For example, in figure 2, if we consider the grey area as the input space associated to class 1 and the white area as the input space associated to class 0, class 1 has 100% of its original input space changed, while class 0 has $1/3 \approx 33\%$ of its associated input space changed. So, the maximum percentage is 100%.

Although class severity can reflect mainly changes in the prior and posterior probabilities, it does not reflect well changes in the unconditional and class-conditional pdfs. In order to cover these changes, we introduced the feature severity criterion, which represents the amount

of changes in the distribution of the input attributes. In the same way as with class severity, drifts can be severe if the probabilities associated to the whole input space are modified and intersected if part of the input space maintains the same probability. This criterion can be measured, for example, by calculating the area between the curves of the old and new unconditional pdfs. Another possible measure, which may be easier to calculate, but less descriptive, would be the percentage of the input space which has its probability modified.

The criterion *speed* has been previously used in the literature, although dividing drifts into vague and heterogeneous categories. Speed is the inverse of the drifting time, which can be measured as the number of time steps taken for a new concept to completely replace the old one [1]. In that way, a higher speed is related to a lower number of time steps and a lower speed is related to a higher number of time steps. According to speed, drifts can be categorized as either abrupt, when the complete change occurs in only 1 time step, or gradual, otherwise.

As explained in [31], sometimes concepts will change gradually, creating a period of uncertainty between stable states. The new concept only gradually takes over and some examples may still be classified according to the old concept. An example given by the author is the behaviour of a device beginning to malfunction – it first fails (classifies in a new way) only sometimes, until the new failure mode becomes dominant. We will refer to this type of drift as probabilistic gradual drift.

Another possibility, not considered in [31], but used in a data set created in [32], is that the concept itself gradually and continuously changes from the old to the new concept, by suffering modifications between every consecutive time step. We will refer to this type of drift as continuous gradual drift. We use the word *continuous* here to refer to changes in which the old concept suffers modifications at **every** time step since the drift started until the new concept is obtained.

As suggested by [31], the speed of a drift can be modelled by a function which represents the degree of dominance of a concept over the other. This idea was further adopted by other authors [13], [25]. Figure 3 shows an example of a drift which takes 100 time steps. In this example, the concept changes linearly and gradually from the old to the new one.

We will consider here that the speed of a drift can also be modelled by a function representing the changes in the old concept, allowing the representation of continuous gradual drifts. For example, consider a moving hyperplane problem:

$$\sum_i a_i x_i \leq a_0 \ .$$

A continuous change from the old to the new concept could be represented by the function:

1. A complete replacement of the old concept means that the examples are generated according to the new concept, and yet the old and new concepts can be intersected.

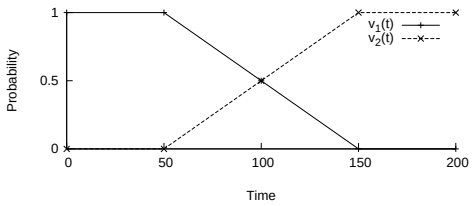

Fig. 3. Example of a gradual drift with drifting time of 100 time steps. The functions $v_1(t)$ and $v_2(t)$ represent the probability that an example from the old and new concept, respectively, will be presented.

$$a_i(t) = a_i(t-1) * 0.001, \ t_d \leq t \leq t_d + drifting\_time \ ,$$

where t_d is the time step in which the drift starts to happen and $drifting\_time$ is the number of time steps for the new concept to completely replace the old concept.

3.2 Criteria to Categorize Drift Sequences

The criterion predictability has not been previously used in the literature either. According to it, drifts can be divided into random and predictable. An example of a predictable drift is a concept represented by a circle whose radius remains fixed and whose centre moves in a straight line.

Recurrence is a criterion that has been previously used in the literature [25]. It divides drifts into recurrent, if there are returns to previous concepts, and non-recurrent, otherwise. Recurrent drifts can have cyclic or unordered (non cyclic) behaviour. An example of an environment with cyclic concepts would be weather forecast, in which different seasons may cause drifts in a cyclic way. An example of an unordered recurrent concept would be the market basket analysis problem, in which a concept drift can be caused by the introduction of a new product. However, after a certain amount of time, the customers can find that the new product is not so good and the concept can return to the previous one. It is interesting to notice, though, that there may be a return to a concept which is similar to a previous one, but not the same. So, a return to a previous concept may be associated to a certain severity in relation to the previously existing concept.

Frequency is another criterion that has not been used in the literature yet. To measure frequency, it is possible to use the number of time steps between the start of 2 consecutive drifts. According to this criterion, drifts can be divided into periodic and non-periodic. If a new concept drift starts to take place at every t time steps, the drifts are considered periodic. Otherwise, they are considered non-periodic. It is interesting to notice that, although recurrence and frequency are distinct criteria, several drifts that present a recurrent behaviour are also periodic. For example, in weather forecast, drifts can happen periodically according to the time of the year, returning to previous concepts depending on the season.

3.3 Remarks on Intermediate Concepts and Additional Comments

Besides defining the criteria and categories, in order to provide a proper categorization, it is necessary to consider if a concept could be denominated as an intermediate concept between the old and the new concept or not. As briefly explained before, consider the concepts represented in figure 1(b). Some authors would regard the concepts represented by the circles between A and B as intermediate concepts between the old concept A and the new concept B.

Depending on the definition of intermediate concept, both the category to which a particular drift belongs and the characterization of the drift according to a particular criterion can change. For example, if we consider I1-3 as intermediate concepts in figure 1(b), there is only one drift (from A to B), which has 100% of severity according to the second measure of severity (M.2). However, if we do not consider I1-3 as intermediate concepts, we would have 4 intersected (not so severe) drifts. If we consider I1-I3 as intermediate concepts active for 100 time steps each and the complete change from a concept to another takes only 1 time step, the drift would be considered gradual. However, if I1-3 are not considered intermediate concepts, we would have 4 abrupt drifts.

A possible definition for intermediate concept could be a concept which is not active for enough time to be PAC [33] learnt. In [34], an upper bound for the maximum number of time steps required for a new concept to be PAC learnt, considering that the previous concept was PAC learnt, is defined. However, such upper bound depends on the VC-dimension [35] of the learning algorithm and a definition of intermediate concept based on this bound would depend on the learning algorithm, not having straightforward application.

Another option for defining intermediate concept would be to consider it as a subjective idea. In this way, before using a particular data set which contains drifts, it would be necessary to subjectively consider a concept as intermediate or not. However, a subjective definition would affect the categorization of drifts and the same category could have very different characteristics depending on the characteristics of the intermediate concepts.

So, in order to provide a simpler and more straightforward categorization, the use of the term intermediate to refer to concepts is not permitted when using the proposed categorization. Every concept should be considered as either old or new and the notion of old and new concepts should only be applied for a particular drift in isolation. We can eliminate the use of the term intermediate concept thanks to the inclusion of criteria not previously used in the literature, such as predictability and severity, which allow drifts to be well described and differentiated even without using this term. By using the proposed criteria, we create a more systematic, well-defined, non-heterogeneous and mutually exclusive categorization than the currently used in the literature, which considers only the criteria speed and recurrence.

It is important to notice that, in order to describe drifts, all the proposed criteria should be considered at the same time. For example, a particular drift can be intersected, abrupt and in a sequence of periodic, random and non-recurrent drifts. Real world problems may present more than 1 drift and each drift can be of a different type.

4 Previous Work on Diversity and Importance of a Diversity Study in Concept Drift

In off-line mode, diversity among base learners is an issue that has been receiving lots of attention in the ensemble learning literature. The success of ensemble learning algorithms is believed to depend both on the accuracy and on the diversity among the base learners [23] and some empirical studies revealed that there is a positive correlation between accuracy of the ensemble and diversity among its members [24], [36]. Breiman's study [37] also shows that lower generalization error random forests have lower base learners correlation and higher base learners strength. Besides, Breiman [37] derives an upper bound for the generalization error of random forests which depends on the correlation and strength of the base learners.

In regression tasks, the bias-variance-covariance decomposition [38] can provide a solid quantication of diversity for linearly weighted ensembles. The decomposition shows that the mean squared error of an ensemble depends critically on the amount of correlation between networks, quantified in the covariance term and that, ideally, we would like to decrease the covariance at the same time as being careful not to increase the bias and the variance terms.

However, there is still no clear analogue of the bias-variance-covariance decomposition when the predictors output discrete labels, as the concept of covariance is not defined. Although there are some theoretical results, they are highly restricted and make strong assumptions unlikely to be hold in practice [39]. It is still not clear how to define a diversity measure for classifiers and how to use it to improve the ensemble's accuracy. In [24], 10 different diversity measures for classifiers are analysed. The authors empirically show that all these measures are correlated to each other, although some of them can exhibit a different behaviour from the others.

In [40], the relationship between diversity and margins of an ensemble is studied considering 6 different diversity measures for classifiers. The authors show that, when the average classification accuracy of the ensemble members on the training data is considered a constant and the maximum diversity is achievable, maximizing the diversity is equivalent to maximizing the minimum margin of the ensemble on the training examples. So, seeking diversity might be viewed as an implicit way

to maximize the minimum margin of the ensemble. However, they theoretically and empirically show that the maximum diversity is usually not achievable. Besides, the minimum margin of an ensemble does not monotonically increase with respect to diversity. Hence, enlarging diversity is not exactly the same as enlarging the minimum margin. Based on that, they conclude that large diversity may not always correspond to better generalization performance.

Furthermore, it is usually affirmed in the literature that there is a trade-off between accuracy and diversity, meaning that lower accuracy may correspond to higher diversity. However, it is shown in [40] that the relationship between accuracy and diversity is not straightforward and lower classification accuracy of ensemble members may not correspond to a higher diversity.

As we can see, the relationship among diversity, accuracy on the training set and generalization is complex. Nevertheless, it is important to study the effect of diversity not only in off-line, but also in on-line changing environments, as the effect of diversity can be very different in these 2 cases.

Although ensembles have been used to handle concept drift as briefly explained in section 1, the literature does not contain any deep study of why they can be helpful for that and which of their features can contribute or not to deal with concept drift. Diversity could be expected to be one of the features that help in dealing with concept drifts when using ensembles. However, there has been no study of the influence of diversity in on-line ensemble learning in the presence of concept drift.

So, it is important to conduct a study checking (1) the differences between the influence of diversity on the ensemble error before and after the drift, (2) whether different types of drift require different amounts of diversity, (3) the influence of diversity on the ensemble's sensitivity to drifts and on the adaptation to the new concept considering base learners that learnt the old concept and (4) whether it is possible to exploit diversity to better handle concept drift.

5 Impact of Diversity on On-line Ensemble Learning

This section presents a diversity study in the presence of concept drift, aiming at analysing all the 4 points commented in the end of section 4.

The rest of this section is organized as follows. Section 5.1 explains the on-line ensemble approach used in the experiments. Section 5.2 explains the data sets used in the experiments and the importance of using artificial data sets when working with concept drift. Section 5.3 explains the experimental design and measures used to get answers to the 4 points commented in the end of section 4. Section 5.4 presents the analysis performed with the results of the experiments.

Inputs: ensemble **h**, ensemble size M, training example d and on-line learning algorithm for the ensemble members $OnlineBaseLearningAlg$.

1: **for** $m = 1$ to M **do**
2: $K \leftarrow Poisson(1)$
3: **while** $K > 0$ **do**
4: $h_m = OnlineBaseLearningAlg(h_m, d)$
5: $K = K - 1$
6: **end while**
7: **end for**

Output: updated ensemble **h**.

Fig. 4. On-line Bagging Algorithm

5.1 On-line Ensemble Approach

The on-line ensemble learning approach used in our experiments is called On-line Bagging [1]. We decided to use this approach because it successfully approximates the well known off-line ensemble learning approach Bagging [41] and does not present any specific behaviour to handle concept drift. It is important to study diversity using an approach that is not specifically designed to deal with concept drift in order not to influence the analysis by other possible mechanisms to handle concept drift. In this way, it is possible to determine the role of diversity by itself in the presence of drifts and whether it is possible to exploit it to better deal with them.

Moreover, we opted for on-line bagging, instead of other approaches such as on-line versions of boosting [1] or random forests [5], because a simple modification including an additional parameter (λ) allows us to tune diversity and this parameter is the only source of diversity apart from the data set itself. So, it is possible to easily increase or reduce diversity by varying this parameter. Section 5.4.1 shows that indeed changing λ consistently changes diversity. Besides, non-weighted ensemble approaches are more appropriate in the context of concept drift [10], when the test data may have a different concept from the most recent training data.

On-line bagging is based on the fact that, when the number of training examples tends to infinite in off-line bagging, each base learner h_m contains K copies of each original training example d, where the distribution of K tends to a $Poisson(1)$ distribution. So, in on-line bagging (figure 4), whenever a training example is available, it is presented K times for each ensemble member h_m, where K is drawn from a $Poisson(1)$ distribution. In our experiments, we slightly modify algorithm 4 to encourage more or less diversity by tuning the parameter λ of the $Poisson(\lambda)$ distribution. The classification is done by unweighted majority vote, as in Bagging.

5.2 Data Sets

When working with real world data sets, it is not possible to know exactly when a drift starts to occur, which type of drift is present or even if there really is a drift. So, it is not possible to perform a detailed analysis of the

behaviour of algorithms in the presence of concept drift using only pure real world data sets. In order to analyse the strong and weak points of a particular algorithm, it is necessary first to check its behaviour using artificial data sets containing simulated drifts. Depending on the type of drift in which the algorithm is weak, it may be necessary to adopt a different strategy to improve it, so that its performance is better when applied to real world problems. In the same way, when analysing strategies that may be adopted in an algorithm to handle concept drift, it is necessary to know in which situations and how these strategies could contribute, i.e., it is necessary to know with which sort of drifts they could help.

In order to analyse the effect of diversity in the presence of concept drift, we developed a data sets generator which can create not only data sets for the 2 most popular concept drift benchmarks (STAGGER boolean concepts [42] and SEA moving hyperplane concepts [11]), but also to different types of drift for 4 problems:

- Circle $(x-a)^2+(y-b)^2 \leq/> r^2$
- Sine $y \ \leq/> \ a\,sin(bx+c)+d$
- Moving hyperplane $y \ \leq/> -a_0+\sum_{i=1}^{d} a_i x_i$
- Boolean
  $y = (color \ =_1/\neq_1 \ a \ \vee_1/\wedge_1$
  $shape \ =_2/\neq_2 \ b) \ \vee_2/\wedge_2$
  $size \ =_3/\neq_3 \ c$

where a, b, c, d, r, a_i, $=/\neq$, $\vee/\wedge$ and $\leq/>$ can assume different values to define different concepts. Besides, the generator allows graphical visualization of the drifts and the generated data sets. The implemented code uses Matlab and is available for download at www.cs.bham.ac.uk/~flm/opensource/DriftGenerator.zip. The examples generated contain x/x_i and y as the input attributes and the concept (which can assume value 0 or 1) as the output attribute.

In this work, we generated data sets for the problems circle, 2 different sines, line (moving hyperplane with $d = 1$), plane (moving hyperplane with $d = 2$) and boolean. Eight irrelevant attributes and 10% class noise were introduced in the plane data sets. Each data set contains 1 drift and different drifts were simulated by varying among 3 amounts of severity (as shown in table 2) and 3 speeds, thus generating 9 different drifts for each problem. The feature severity affects the same areas of the input space as the class severity for all the problems but plane and boolean. So, we will refer to class severity simply as severity in our experiments. For plane and boolean, there is no feature change. The speed was modelled by the following linear degree of dominance functions:

$$v_n(t) = \frac{t-N}{drifting\_time}, \ N < t \leq N + drifting\_time$$

and

$$v_o(t) = 1 - v_n(t), \ N < t \leq N + drifting\_time \ ,$$

where $v_n(t)$ and $v_o(t)$ are the degrees of dominance of the new and old concepts, respectively; t is the current time step; N is the number of time steps before the drift started to occur; and $drifting\_time$ varied among 1, $0.25N$ and $0.50N$ time steps.

TABLE 2
Artificial Data Sets

Probl.	Fixed Values	Before→After Drift	Class Severity M.1
Circle	$a=b=0.5$ $\leq$	$r=0.2 \rightarrow 0.3$	$\approx 16\%$
		$r=0.2 \rightarrow 0.4$	$\approx 38\%$
		$r=0.2 \rightarrow 0.5$	$\approx 66\%$
SineV	$a=b=1$ $c=0$ $\leq$	$d=-2 \rightarrow 1$	15%
		$d=-5 \rightarrow 4$	45%
		$d=-8 \rightarrow 7$	75%
SineH	$a=d=5$ $b=1$ $\leq$	$c=0 \rightarrow -\pi/4$	$\approx 36\%$
		$c=0 \rightarrow -\pi/2$	$\approx 57\%$
		$c=0 \rightarrow -\pi$	$\approx 80\%$
Line	$a_1=0.1$ $\leq$	$a_0=-0.4 \rightarrow -0.55$	15%
		$a_0=-0.25 \rightarrow -0.7$	45%
		$a_0=-0.1 \rightarrow -0.8$	70%
Plane	$a_1=a_2=0.1$ $\leq$	$a_0=-2 \rightarrow -2.7$	14%
		$a_0=-1 \rightarrow -3.2$	44%
		$a_0=-0.7 \rightarrow -4.4$	74%
Bool	$c=S \vee M \vee L$	$a=R, \wedge_1$ $b=R \rightarrow R \vee T$	$\approx 11\%$
	$\wedge_2$	$a=R, b=R,$ $\wedge_1 \rightarrow \vee_1$	$\approx 44\%$
	$=_1=_2=_3$	$a=R \rightarrow R \vee G,$ $b=R \rightarrow R \vee T,$ $\wedge_1 \rightarrow \vee_1$	$\approx 67\%$

The training sets are composed of $2N$ examples and each example corresponds to 1 time step of the learning. The first N examples of the training sets were generated according to the old concept ($v_o(t) = 1$, $1 \leq t \leq N$), where $N = 1000$ for circle, sineV, sineH and line and $N = 500$ for plane and boolean. The next $drifting\_time$ training examples ($N < t \leq N + drifting\_time$) were generated according to the degree of dominance functions, $v_n(t)$ and $v_o(t)$. The remaining examples were generated according to the new concept ($v_n(t) = 1$, $N + drifting\_time < t \leq 2N$).

Several authors use different ways to test their algorithms in on-line and incremental learning. Some authors create a test set which reflects exactly the underlying distribution of the train data at the time step to be tested [11], [18]. Other authors use the on-line error, calculated by updating the average with the prediction of each example before its learning [13], [14]. Others consider that the test error should not reflect exactly the underlying distribution of the train data, as, in the real world, the distribution of the problem may change since the presentation of the last training examples [10]. Besides, in incremental learning, the algorithms are frequently tested with the next chunk of data to be learnt, before its learning [30], [43].

We consider that test data should be created using different distributions from train data especially when the drifts are gradual. So, in this paper, we create data sets divided into 2 partitions. The first partition is composed of $0.25N$ examples of the old concept and the second is composed of $0.25N$ examples of the new

concept. In order to test the system before the drift, only the examples belonging to the first concept are used, whereas to test the system after the begin of the drift, only the examples of the new concept are used, even if the drift is still not completed. In this way, it is possible to check the behaviour of algorithms in relation to the old and the new concept and simulate the situation in which the distribution of the test set is different from the distribution of the train set for gradual drifts.

The range of x or x_i was [0,1] for circle, line and plane; [0,10] for sineV; and [0,4*π] for sineH. The range of y was [0,1] for circle and line, [-10,10] for sineV, [0,10] for sineH and [0,5] for plane. For plane and boolean, the input attributes are normally distributed through the whole input space. For the other problems, the number of instances belonging to class 1 and 0 generated for both the train and the test data set is always the same, having the effect of changing the unconditional pdf when the drift occurs.

Besides the data sets created using the data sets generator, we also created partially artificial data sets based on the real world databases Contraceptive, Yeast, Iris and Car, from the UCI Machine Learning Repository [44], by introducing simulated drifts inspired by [7]. The car database was randomized and divided into 3 partitions in order to simulate 2 drifts (the first partition represents the examples before any drift; the second partition represents examples after the begin of the first drift and before the second drift; and the third partition represents examples after the begin of the second drift). The other databases were replicated 3 (or 4, for yeast) times in order to create the 3 (or 4) partitions and each partition was randomized. Different concepts were created by changing the labels of the target classes of the examples according to table 3. As it can be observed from the table, all the second drifts present a partial return to the first concept, i.e., part of the changes caused by the first drift are undone by the second drift.

Each partition has size N ($N = 1473$ for contraceptive, 1482 for yeast, 150 for iris and 576 for car) and was further divided into train (75%) and test (25%) examples. The $drifting\_time$ for the first drift was $0.25N$ for contraceptive and yeast and 1 for iris and car. The $drifting\_time$ of the second (and third) drift was 1 for contraceptive, $0.25N$ for yeast and $0.5N$ for iris and car. The speed is modelled by linear degree of dominance functions, similarly to the data sets created using the data sets generator. The test examples used in each time step are also chosen in the same way as for those data sets, representing the first concept when no drift started to happen and representing only the new concept after the begin of a drift.

TABLE 3
UCI Data Sets - Rounded percentage of examples of each class in the original database and concepts (C_i) used to create the drifting data sets.

Probl.	Original class	C1	C2	C3	C4
Cont.	No-use 42.70%	1	2	2	-
	Long-term 22.61%	2	3	1	-
	Short-term 34.69%	3	1	3	-
Yeast	CYT 31.16%	1	2	1	1
	NUC 28.87%	2	1	2	2
	MIT 16.42%	3	4	4	4
	ME3 10.97%	4	3	3	3
	ME2 3.43%	5	5	5	6
	ME1 2.96%	6	6	6	5
	EXC 2.49%	7	7	7	8
	VAC 2.02%	8	8	8	7
	POX 1.35%	9	9	9	10
	ERL 3.36%	10	10	10	9
Iris	Setosa 33.33%	1	4	4	-
	Versicolour 33.33%	2	1	2	-
	Virginica 33.33%	3	3	3	-
	0%	4	2	1	-
Car	Unacc 70.02%	0	1	0	-
	Acc 22.22%	1	0	1	-
	Good 3.99%	1	0	0	-
	Very good 3.76%	1	0	0	-

5.3 Experimental Design and Measures Analysed

The experiments concentrate on answering points (1)-(4) commented in the end of section 4. In order to do so, we first perform analysis of variance (ANOVA) [45] to analyse the impact of λ on diversity when using on-line bagging (section 5.4.1) based on the artificial data sets. Then, we perform ANOVA to analyse the impact of λ and other factors on the test error to get answers to the points (1) and (2) (section 5.4.2). After that, we analyse the test error over time in order to check how sensitive to drifts ensembles with different diversity levels are and how fast they recover from drifts (section 5.4.3), getting answers to points (3) and (4). Finally, we present the results obtained using the UCI data sets, reassuring the analysis (section 5.4.4).

The experiments with the artificial problems use a split-plot (mixed) design, where a between-subjects design is combined with a repeated measures design, whereas the experiments with the UCI problems use a repeated measures design [45].

The within-subject factors are the parameter λ from on-line bagging and the time step analysed. The factor λ varied among 8 different levels: 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1 and was used to encourage different amounts of diversity. The factor time step varied among $0.99N$, $1.1N$, $1.5N$ and $2N$ for the artificial data sets and also included $2.1N$, $2.5N$, $3N$ (and $3.1N$, $3.5N$ and $4N$ for yeast) for the UCI data sets. These time steps were chosen in order to analyse the response shortly before the drift, shortly after the drift and longer after the drift.

The between-subject factors used in the split-plot design are severity and speed, which varied among 3 different levels each and were used to create the artificial data sets, as explained in section 5.2.

When analysing the impact of λ on diversity, the response is the Q statistics [24], which is one of the most often used diversity measures for classifiers. Q

statistics can assume values in [-1,1] and lower values represent more diversity. When analysing the impact of λ on the test error, the response at the time step $0.99N$ is the classification error on the fraction of the test set corresponding to the old concept and the response at the other time steps is the classification error on the fraction of the test set corresponding to the new concepts. So, the response represents how well the old/new concepts were learnt.

Thirty repetitions for each combination of factor level (except time) were done, totalizing 2160 executions for each artificial problem and 240 executions for each UCI problem. The ensembles were composed of 25 ITI on-line decision trees [46].

5.4 Experimental Results and Analysis

This section presents the analysis performed with the results of the experiments. Section 5.4.1 explains how λ influences diversity when using on-line bagging. Section 5.4.2 concentrates on answering points (1) and (2) mentioned in the end of section 4. Section 5.4.3 concentrates on answering points (3) and (4). Finally, section 5.4.4 reassures the analysis, by presenting the results obtained with the UCI data sets.

5.4.1 The Impact of λ on Diversity

As explained in section 5.1, we use different λ values to encourage different amounts of diversity. In this section, we explain that higher λ values are in general associated to lower diversity average and lower λ values are associated to higher diversity average. In order to do so, we performed ANOVA to analyse the influence of each factor mentioned in section 5.3 on the response, which is in this case the Q statistics [24], for the artificial data sets. Mauchly's tests of sphericity [47] detected violations of the sphericity assumption (null hypothesis always rejected with p-value less than 0.001), so Greenhouse-Geisser corrections [48] were used. The ANOVA results table was omitted due to space limitations.

ANOVA shows that λ has the largest effect size (considering the eta-squared [49], [50] [2]) in the within-subjects tests for all the problems. The effect size is very large – eta-squared always more than 0.90 and usually more than 0.97. The eta-squared values for the other factors and interactions are usually very low (under 0.015). In the between-subjects tests, severity usually has large effect size (eta-squared larger than 0.10 and lower than 0.60) whereas eta-squared associated to speed and the interaction severity*speed are usually lower than 0.010. So, the factors with more influence on the response are λ and severity.

As we will need to analyse the interactions λ*time*severity in the next sections, we created all the plots of marginal means for λ vs. severity and λ vs. speed for each time step analysed. In this way, it is possible to check the average Q statistics associated to each one of these combinations.

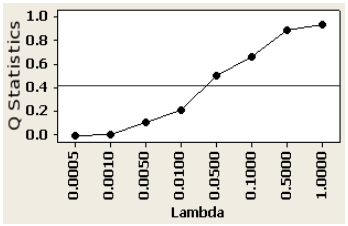

Fig. 5. Plot of the main effect of λ on Q statistics for circle.

The plots show that, for all the cases, higher λ values generate higher Q statistics (lower diversity) and lower λ values generate lower Q statistics (higher diversity). Besides, the lowest Q statistics is always close to 0 and the highest is always close to 1 (although always lower than 1). Frequently, the Q statistics values generated for $\lambda = 0.0005$ and 0.001 are similar to each other. The same happens for $\lambda = 1$ and 0.5. As different amounts of severity or speed usually do not change much the response and never cause a higher λ to have lower Q statistics than a lower λ, we present the plot of the main effect of λ on the Q statistics for circle in figure 5. The plots for the other problems are similar and were omitted due to space limitations.

5.4.2 The Impact of Diversity on the Test Error Before and After a Drift

In this section, we concentrate mainly on checking the points (1) and (2) mentioned in the end of section 4: (1) the differences between the influence of diversity on the ensemble error before and after the drift and (2) whether different types of drift require different amounts of diversity. In order to do so, we performed ANOVA to analyse the influence of each factor mentioned in section 5.3 on the response, which is in this case the test error, as explained in section 5.3. The data sets used in this analysis are the artificial data sets presented in section 5.2. We will consider the influence of λ as the influence of diversity on the response, based on section 5.4.1.

Table 4 shows results of the tests of within-subjects for each problem. The table presents the p-value, the type III sum of squares (SS), degrees of freedom (DF), mean squares (MS), test F statistics (F) and (Eta) eta-squared. Interactions between 2 factors are represented by factor1*factor2. P-values less than 0.01 represent rejection of the null hypothesis that the average response is statistically equal at all the levels of the corresponding factors, considering significance level of 1%. Only factors/interactions involving λ or with large effect size (eta-squared higher than 0.10) are shown in order to reduce the space occupied by the table. Mauchly's tests of sphericity [47] detected violations of the sphericity assumption (p-value always less than 0.001), so Greenhouse-Geisser corrections [48] were used.

2. For a split-plot ANOVA, eta-squareds need to be calculated separately within the context of the within-subject effects ANOVA table and the between-subject effects ANOVA table. In this situation, the eta-squareds will sum to 1 for within-subjects and will sum to 1 for the between-subject effects.

TABLE 4
ANOVA – Test of Within-Subjects Effects: Factors/interactions which involve λ or have eta-squared higher than 0.10, in order of effect size for each problem. The p-value is always less than 0.001, except for λ*Sev*Sp for SineH, in which it was 0.009.

Prob	Factor/Int.	SS	DF	MS	F	Eta	Prob	Factor/Int.	SS	DF	MS	F	Eta
Circle	λ	95.59	2.59	36.85	7130.56	.559	Line	λ	119.56	1.70	70.47	6709.61	.564
	Time	31.03	2.77	11.20	10991.84	.181		Time	39.28	2.11	18.60	15127.60	.185
	λ * Time	22.97	8.07	2.84	1402.40	.134		λ * Time	20.04	4.70	4.26	1150.41	.095
	λ * Sev	5.74	5.19	1.11	214.19	.034		λ * Sev	8.07	3.39	2.38	226.42	.038
	λ * Time * Sev	2.14	16.15	.132	65.27	.012		λ * Time * Sev	3.10	9.40	.33	89.07	.015
	λ * Sp	.29	5.19	.06	10.81	.002		λ * Time * Sp	.15	9.40	.02	4.19	.001
	λ * Time * Sp	.23	16.15	.01	7.01	.001							
SineV	λ	114.27	1.77	64.44	6800.64	.527	Plane	Time	39.28	2.11	18.60	15127.60	.185
	Time	46.03	2.34	19.65	16944.44	.212		λ	58.61	3.85	15.24	933.81	.142
	λ * Time	21.08	5.45	3.87	1191.54	.097		λ * Sev	36.00	7.69	4.68	286.78	.087
	λ * Sev	7.87	3.55	2.22	234.31	.036		λ * Time	23.25	6.24	3.72	241.78	.056
	λ * Time * Sev	3.66	10.90	.33	103.35	.017		λ * Time * Sev	7.37	12.48	.59	38.32	.018
	λ * Sp	.20	3.55	.06	6.09	.001		λ * Sp	.96	7.69	.12	7.62	.002
	λ * Time * Sp	.31	10.90	.03	8.82	.001		λ * Time * Sp	.94	12.48	.07	4.90	.002
SineH	Time	119.82	2.60	46.15	40623.55	.46	Bool	Time	198.23	2.05	96.84	23125.09	.594
	λ	44.32	3.56	12.43	3477.10	.172		Time*Sev	64.09	4.09	15.65	3738.36	.192
	λ * Time	41.65	7.82	5.32	2043.49	.161		λ	29.34	3.70	7.94	918.07	.088
	λ * Sev	14.84	7.13	2.08	582.17	.057		λ * Time	8.17	6.13	1.33	166.48	.024
	λ * Time * Sev	6.06	15.65	.39	148.70	.023		λ * Sev	5.13	7.39	.69	80.21	.015
	λ * Sp	.24	7.13	.03	9.47	.001		λ * Time * Sev	2.47	12.27	.20	25.18	.007
	λ * Time * Sp	.18	15.65	.01	4.42	.001		λ * Time * Sp	.64	12.27	.05	6.49	.002
	λ * Sev * Sp	.11	14.26	.01	2.13	.000							

The interaction between time and severity (usually not reported in the table) has frequently medium and in 1 case large effect size, whereas interactions between speed and other factors is usually small, showing the importance of the drifts categorization proposed in section 3, which distinguishes severity and speed.

As it can be observed from table 4, not only time, but also λ usually has large effect size. This shows that, not only the drift, but also diversity has large impact on the response. Excluding boolean, the interactions between λ and severity always have medium effect size (eta-squared between 0.034 and 0.087) and the interactions between λ and time always have medium or large effect size (eta-squared between 0.056 and 0.161). This shows that diversity plays important and probably different roles depending on the severity and time (before, shortly after or longer after drift). The effect size of the interactions involving speed is always very small (eta-squared 0.002 or less). So, in the rest of this section, we will check what role diversity plays depending on severity and time.

As there are interactions among λ, time and severity for all the problems, we generated plots of marginal means λ vs. severity for each problem and time step analysed. The plots for circle are shown in figure 6. Plots for other problems are always omitted in this section due to space limitations.

According to the plots, at the time step $0.99N$ (before drift), the highest λ (lowest diversity) obtain the best responses independent on the severity. The only exception is plane, which is the only problem that contains many irrelevant attributes. In this case, $\lambda = 0.1$ obtained the best responses. For boolean, $\lambda = 0.05$, 0.1, 0.5 and 1 obtained always response 0 (zero) at this time step. Please, notice that the accuracy of an ensemble depends not only on the diversity, but also on the accuracy of each ensemble member, as explained in section 4. So, only maximizing the diversity is not the same as maximizing the ensemble accuracy and it is reasonable that the lowest λs did not lead to the highest ensemble accuracies.

At the time step $1.1N$ (shortly after the drift), when severity is high, lower λs frequently obtain better responses than $\lambda = 1$. As severity reduces, the best λs start obtaining more similar response to $\lambda = 1$. Besides, the best λ values tend to be higher for the low than for the high severity drifts (higher for 4 problems and the same for 2 problems). So, the higher the severity, the more beneficial is high diversity shortly after the drift.

Additional paired T tests [51] using Bonferroni corrections considering all combinations of severity and diversity with overall significance level of 0.01 (so that a test is considered signicant if its associated probability is smaller than $0.01/24$) confirm this analysis. They show that, when severity is low, the null hypothesis that the average response is statistically equal using the best average response $\lambda < 1$ and $\lambda = 1$ is rejected for only 2 in 6 problems. However, when severity is medium or high, this number increases to 4.

That is an interesting fact, as one could expect that higher severity would make different diversity ensembles equally bad right after the drift, but, actually, highly diverse ensembles manage to get better responses when severity is high and it is the lowest severity which makes

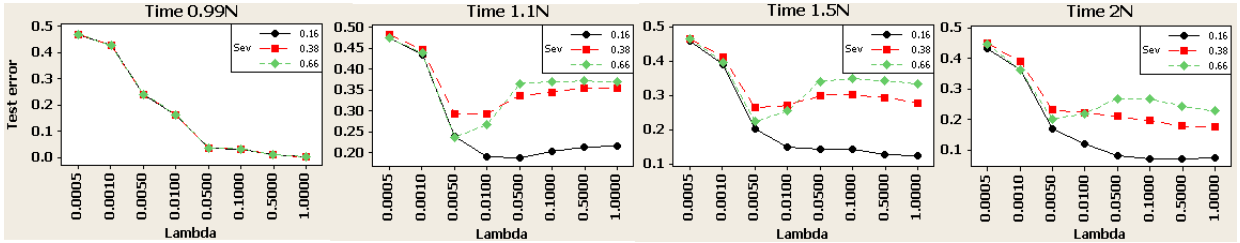

Fig. 6. Plots of marginal means for the effect of λ*severity*time on the test error for circle.

the behaviour of high and low diversity ensembles more similar. Another interesting fact is that, even though high diversity is more beneficial when severity is high, the response of the best $\lambda < 1$ is always better or similar to $\lambda = 1$, independent on the severity.

At the time step $1.5N$, high diversity is still more important when severity is higher. Besides, additional T tests with Bonferroni correction reveal that the number of problems in which the best λ obtains statistically significant different average response from $\lambda = 1$ for low and high severity increases (from 2 to 3 and from 4 to 5, respectively). So, higher diversity is still important at the time step $1.5N$.

At the time step $2N$, the importance of high diversity reduces for all the severities (from 3 to 2, 4 to 2 and 5 to 3 for low, medium and high severities, respectively).

So, we verified that:

1) There are differences between the influence of diversity on the ensemble error before and after the drift. Before the drift, less diversity usually obtains the best results. However, it is a good strategy to maintain high diversity in order to obtain better results after the drift. After a large number of time steps have passed since the beginning of the drift, high diversity becomes less important.
2) Drifts with different severities require different amounts of diversity (higher diversity is more important for more severe drifts). However, the effect of diversity on drifts with different speeds is very small.

Besides these points, it is also good to verify which diversity level generally obtains the best average responses, considering all types of drift at the same time. So, it is important to analyse the main effect plots of λ for each time step and check the Q statistics associated to the best λs. The main effect plots for circle are shown in figure 7. Although there are frequently 2 exceptions (usually boolean and plane), we can make the following observations from the main effect plots. The λ values associated to Q statistics higher than 0.85 usually obtain the best average responses before the drift. However, at the time steps $1.1N$ and $1.5N$, Q statistics lower than 0.25 are usually required. At the time step $2N$, Q statistics higher than 0.95 are frequently among the best.

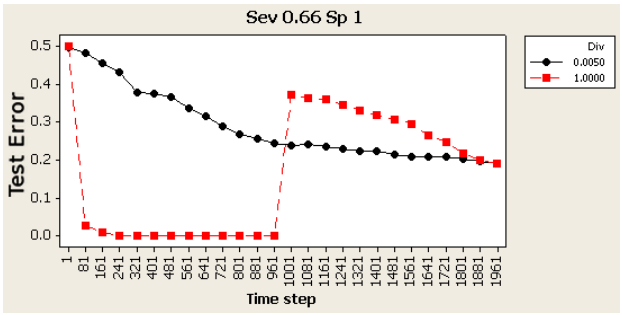

Fig. 8. Average test error for the highest severity and speed drift for circle. The λ value corresponds to the best test error before the drift ($\lambda = 1$) and the best test error shortly after the drift ($\lambda = 0.005$).

5.4.3 *The Influence of Diversity on the Sensitivity to Drifts and Adaptation to the New Concept*

This section concentrates on checking the points (3) and (4) presented in the end of section 4: (3) the influence of diversity on the ensemble's sensitivity to drifts and on the adaptation to the new concept considering base learners that learnt the old concept and (4) whether it is possible to exploit diversity to better handle drifts.

First we shall concentrate on point (3). In section 5.4.2, we checked that high diversity ensembles are desirable soon after the drift, for getting similar or better test error than low diversity ensembles. However, after a large number of time steps have passed since the beginning of the drift, high diversity becomes less important. This is an indicator that, although high diversity ensembles may help to reduce the initial increase in the error soon after the drift (sensitivity to drifts), they are likely not to adapt quickly to the new concept (recovery from drifts).

In order to check if that really happens, we plotted the average test error over time for the λ which obtained the best test error before the drift (time step $0.99N$) and the λ which obtained the best test error soon after the drift (time step $1.1N$). As high diversity helps especially drifts with high severity, we plotted these graphics for the experiments with high severity and high speed. The plots for circle are shown in figure 8.

The plots show that the average test error for the higher diversity ensembles is lower shortly after the drift for all problems but boolean, in which it is similar. However, the average test error for the lower diversity ensembles always decays faster, so that, in the end of the

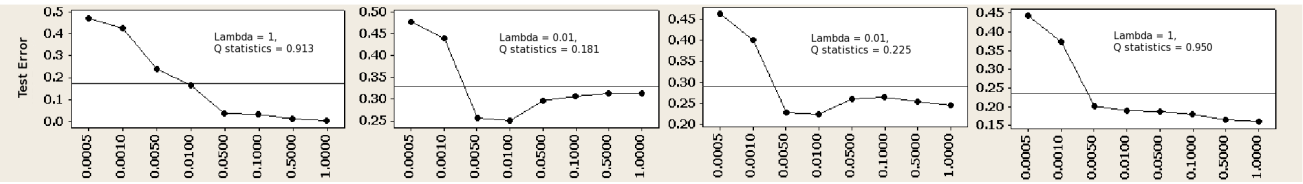

Fig. 7. Plots of the main effect of λ on the test error at the time steps $0.99N$, $1.1N$, $1.5N$ and $2N$ for circle. Q statistics corresponding to the λs with the lowest test errors are indicated.

learning, it usually gets similar or lower test error to the higher diversity ensembles. So, high diversity by itself cannot help the ensemble to have a faster recovery from the drift after the initial effect of reducing the increase in the error. Low diversity can get lower error in the absence of drifts and get faster recovery from drifts, although it still takes a too long time to attain low error after a drift.

Now, we can address point (4). When designing an approach to handle concept drift, 3 issues should be considered. One of them is the speed of recovering (adaptation to the new concept), which is the most addressed issue in the literature. The other one is the reduction of the initial increase in error which occurs right after a drift (sensitivity to drifts). The third one is low error in the absence of drifts. In order to successfully handle concept drift, an approach should address all these issues. So, the answer to point (4) is that diversity can be exploited to better handle concept drift. A well-designed approach which maintains more than one ensemble with different diversity levels may be able to converge in the absence of drifts and get lower test error soon after a drift. However, additional procedures should be adopted in order to properly converge to a new concept, for example, by creating a new ensemble after a drift detection [12]–[15].

5.4.4 Analysis Using UCI Problems

This section presents the results of the experiments performed with the UCI problems. ANOVA indicates that there is interaction between λ and time for all the problems (null hypothesis of equal means always rejected with p-value less than 0.001). The plots of marginal means are shown in figure 9. Only λs corresponding to the best test errors after the drifts and $\lambda = 1$ are shown in order to facilitate reading. We have shown in section 5.4.2 that different severities require different amounts of diversity. Here, the best $\lambda < 1$ also varies for different drifts.

As it can be seen, although $\lambda = 1$ obtains good test errors before the drifts, it gets worse than a lower λ (higher diversity) after the first drift for all the problems. For iris, there is a delay and a lower λ gets lower test error than $\lambda = 1$ only after the time step $1.1N$, but, even so, it gets lower test error. We can also observe that very low λs (e.g., 0.0005 for car and contraceptive) do not converge. Not so low λs present an error decay more similar to $\lambda = 1$. However, they still do not attain low test error on the new concept.

As the second drift presents a partial return to the first concept, higher λs (lower diversity) than after the first drift present lower test error. However, except for yeast, λs lower than 1 still present the lowest errors. A possible explanation for the fact that $\lambda = 1$ obtained good test error after the second and third drifts for yeast is that the new concepts are actually easier than the old concepts. Such situation can happen in real world problems.

The experiments using UCI real world problems with simulated drifts reassure the results presented in the previous sections, showing that high diversity is useful to reduce the initial increase in error when drifts happen. Besides, these experiments suggest that it is worth maintaining ensembles which learnt old concepts, so that they can be used in the case of recurrent drifts.

6 Conclusions

This paper presents a new categorization for concept drift, which separates drifts according to different criteria into mutually exclusive and non-heterogeneous categories. This categorization allows more systematic and detailed studies of drifts. Different aspects of a drift have different impacts on the test error, showing the importance of differentiating them in a drifts categorization.

A diversity analysis using concept drift problems is also presented. The analysis shows that diversity plays an important role both before and after a concept drift. Answering points (1) and (2) commented in section 4, before the drift, ensembles with less diversity obtain better test errors. On the other hand, shortly after the drift, more diverse ensembles are always among the best test errors and the difference in the test error in comparison to lower diversity ensembles is usually more significant when severity is higher. It is a good strategy to maintain highly diverse ensembles to obtain good responses shortly after the drift, independent on the type of the drift. At the time step $2N$, less diversity frequently returns to obtain good responses, although some drifts can still require higher diversity.

Diversity can help mainly to reduce the initial increase in the error caused by a drift (points (3) and (4)). However, some other mechanism is still necessary for the ensemble to recover from the drift and converge to the new concept. A well-designed approach which maintains more than one ensemble with different diversity levels may be able to converge in the absence of drifts and get lower test error soon after a drift. However,

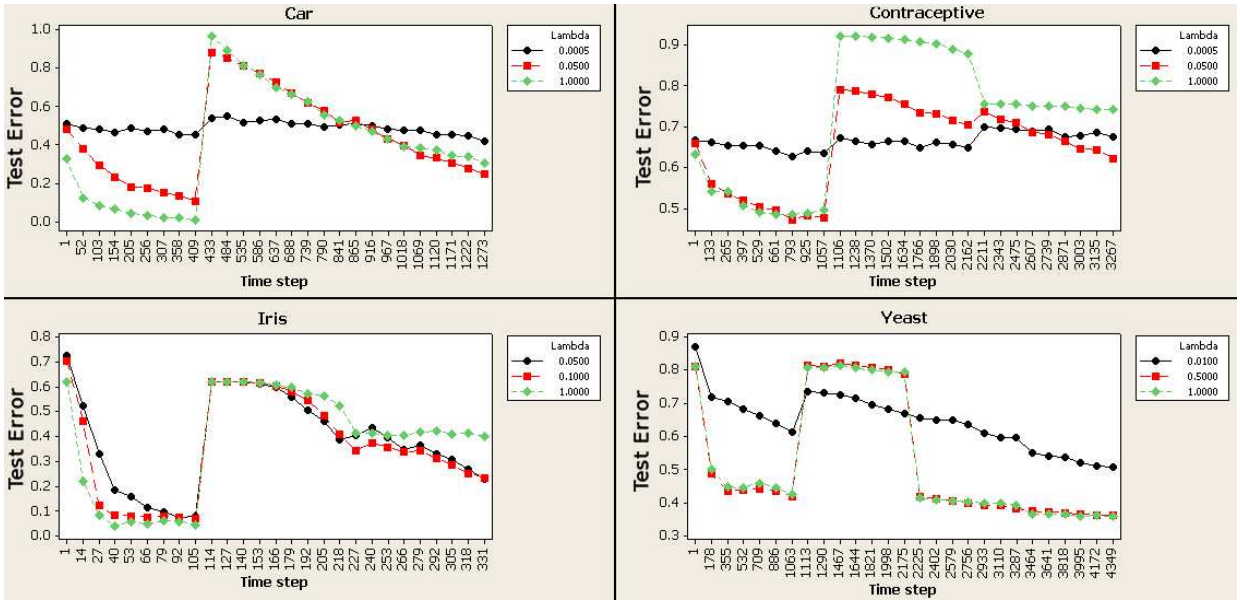

Fig. 9. Average test error for λs corresponding to the best test errors after the drifts and $\lambda = 1$, for the UCI problems.

additional procedures should be adopted in order to properly converge to a new concept, for example, by creating a new ensemble after a drift detection.

Future works include further study of diversity in drift sequences and recurrent drifts and the analysis of a proposed approach to investigate more the point (4). Besides, the impact of the ensemble size and the error obtained by each ensemble member should also be investigated. The difficulty of the concept which was learnt before a drift could also influence the learning of a new concept. We noticed in our experiments that a concept easily learnt might increase considerably more the error obtained after a drift and the concept might be more difficult to be forgotten, depending on the difficulty of the new concept. So, further studies should be done to investigate that. Another hypothesis to be investigated is that higher diversity may help learning when there are many irrelevant attributes.

Acknowledgements

The authors are grateful to Dr. Nikunj C. Oza for sharing his implementation of On-line Bagging.

References

[1] N. C. Oza and S. Russell, "Experimental comparisons of online and batch versions of bagging and boosting," in *Proc. of ACM SIGKDD*, San Francisco, 2001, pp. 359–364.

[2] A. Fern and R. Givan, "Online ensemble learning: An empirical study," *Machine Learning*, vol. 53, pp. 71–109, 2003.

[3] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: An incremental learning algorithm for supervised neural networks," *IEEE Transactions on Systems, Man, and Cybernetics - Part C*, vol. 31, no. 4, pp. 497–508, 2001.

[4] F. L. Minku, H. Inoue, and X. Yao, "Negative correlation in incremental learning," *Natural Computing Journal - Special Issue on Nature-inspired Learning and Adaptive Systems*, p. 32p, 2008.

[5] H. Abdulsalam, D. B. Skillicorn, and P. Martin, "Streaming random forests," in *Proc. of IDEAS*, Banff, Canada, 2007, pp. 225–232.

[6] H. Wang, W. Fan, P. S. Yu, and J. Han, "Mining concept-drifting data streams using ensemble classifiers," in *Proc. of ACM KDD*. New York: ACM Press, 2003, pp. 226–235.

[7] M. Scholz and R. Klinkenberg, "An ensemble classifier for drifting concepts," in *Proc. of the Second International Workshop on Knowledge Discovery from Data Streams*, Porto, 2005, pp. 53–64.

[8] ——, "Boosting classifiers for drifting concepts," *IDA - Special Issue on Knowledge Discovery From Data Streams*, vol. 11, no. 1, pp. 3–28, 2007.

[9] H. He and S. Chen, "IMORL: Incremental multiple-object recognition and localization," *IEEE Transactions on Neural Networks*, vol. 19, pp. 1727–1738, 2008.

[10] J. Gao, W. Fan, and J. Han, "On appropriate assumptions to mine data streams: Analysis and practice," in *Proc. of IEEE ICDM*, Omaha, NE, 2007, pp. 143–152.

[11] W. Street and Y. Kim, "A streaming ensemble algorithm (SEA) for large-scale classification," in *Proc. of ACM KDD*, 2001, pp. 377–382.

[12] F. Chu and C. Zaniolo, "Fast and light boosting for adaptive mining of data streams," in *Proc. of PAKDD'04*, Sydney, 2004, pp. 282–292.

[13] M. Baena-García, J. Del Campo-Ávila, R. Fidalgo, and A. Bifet, "Early drift detection method," in *Proc. of the 4th ECML PKDD International Workshop on Knowledge Discovery From Data Streams (IWKDDS'06)*, Berlin, Germany, 2006, pp. 77–86.

[14] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with drift detection," in *Proc. of the 7th Brazilian Symposium on Artificial Intelligence (SBIA'04) - Lecture Notes in Computer Science*, vol. 3171. São Luiz do Maranhão, Brazil: Springer, 2004, pp. 286–295.

[15] K. Nishida and K. Yamauchi, "Detecting concept drift using statistical testing," in *Proceedings of the Tenth International Conference on Discovery Science (DS'07) - Lecture Notes in Artificial Intelligence*, vol. 3316, Sendai, Japan, 2007, pp. 264–269.

[16] S. Ramamurthy and R. Bhatnagar, "Tracking recurrent concept drift in streaming data using ensemble classifiers," in *Proc. of ICMLA'07*, Cincinnati, Ohio, 2007, pp. 404–409.

[17] G. Forman, "Tackling concept drift by temporal inductive transfer," in *Proc. of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, 2006, pp. 252–259.

[18] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: A new ensemble method for tracking concept drift," in *Proceedings of ICDM*, USA, 2003, pp. 123–130.

[19] ——, "Using additive expert ensembles to cope with concept drift," in *Proc. of ICML'05*, Bonn, Germany, 2005, pp. 449–456.

[20] W. Fan, "Streamminer: a classifier ensemble-based engine to mine concept-drifting data streams," in *Proc. of the 30th International Conference on Very Large Data Bases*, Toronto, 2004, pp. 1257–1260.

[21] ——, "Systematic data selection to mine concept-drifting data streams," in *Proc. of the 10th ACM KDD*, Seattle, 2004, pp. 128–137.
[22] J. Gao, W. Fan, J. Han, and P. Yu, "A general framework for mining concept-drifting data streams with skewed distributions," in *Proc. of SIAM ICDM*, Minneapolis, Minnesota, 2007.
[23] T. G. Dietterich, "Machine learning research: Four current directions," *Artificial Intelligence*, vol. 18, no. 4, pp. 97–136, 1997.
[24] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, pp. 181–207, 2003.
[25] A. Narasimhamurthy and L. I. Kuncheva, "A framework for generating data to simulate changing environments," in *Proceedings of the 25th IASTED AIA*, Innsbruck, Austria, 2007, pp. 384–389.
[26] J. C. Schlimmer and D. Fisher, "A case study of incremental concept induction," in *Proceedings of the 5th AAAI*, Philadelphia, USA, 1986, pp. 496–501.
[27] J. Branke, "Evolutionary algorithms for dynamic optimization problems - a survey," Insitute AIFB, University of Karlsruhe, Tech. Rep. 387, 1999.
[28] ——, *Evolutionary Optimization in Dynamic Environments*. Netherlands: Kluwer Academic Publishers, 2002.
[29] Y. J. and J. Branke, "Evolutionary optimization in uncertain environments - a survey," *IEEE Transactions on Evolutionary Computation*, vol. 9, pp. 303–317, June 2005.
[30] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections," in *Proc. of the 19th IEEE International Symposium on Computer-Based Medical Systems (CBMS'06)*, Salt Lake City, UT, 2006, pp. 56–68.
[31] G. Widmer and M. Kubat, "Learning in the presence of concept drift and hidden context," *Machine Learning*, vol. 23, pp. 69–101, 1996.
[32] K. Nishida and K. Yamauchi, "Adaptive classifiers-ensemble system for tracking concept drift," in *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics (ICMLC'07)*, Honk Kong, 2007, pp. 3607–3612.
[33] L. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.
[34] A. Kuh, T. Petsche, and R. L. Rivest, "Learning time-varying concepts," in *Proc. of NIPS*, San Francisco, CA, 1990, pp. 183–189.
[35] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities." *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971.
[36] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.
[37] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[38] N. Ueda and R. Nakano, "Generalization error of ensemble estimators," in *Proc. of International Conference on Neural Networks*. San Francisco, CA: Morgan Kaufmann Publishers, 1996, p. 9095.
[39] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: A survey and categorisation," *Journal of Information Fusion*, vol. 6, pp. 5–20, 2005.
[40] E. K. Tang, P. N. Sunganthan, and X. Yao, "An analysis of diversity measures," *Machine Learning*, vol. 65, pp. 247–271, 2006.
[41] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.
[42] J. C. Schlimmer and R. H. Granger Jr., "Incremental learning from noisy data," *Machine Learning*, vol. 1, pp. 317–354, 1986.
[43] M. Scholz and R. Klinkenberg, "Boosting classifiers for drifting concepts," *IDA - Special Issue on Knowledge Discovery from Data Streams*, vol. 11, no. 1, pp. 3–28, 2007.
[44] D. Newman, S. Hettich, C. Blake, and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html
[45] D. C. Montgomery, *Design and Analysis of Experiments*, 6th ed. Great Britain: John Wiley and Sons, 2004.
[46] P. Utgoff, N. Berkman, and J. Clouse, "Decision tree induction based on efficient tree restructuring," *Machine Learning*, vol. 29, no. 1, pp. 5–44, 1997.
[47] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *Annals of Mathematical Statistics*, vol. 11, pp. 204–209, 1940.
[48] S. Greenhouse and S. Geisser, "On methods in the analysis of profile data," *Psychometrika*, vol. 24, pp. 95–112, 1954.
[49] D. Howell, *Statistical Methods for Psychology*. Belmont, California: Thomson Wadsworth, 2007.
[50] C. A. Pierce, R. A. Block, and H. Aguinis, "Cautionary note on reporting eta-squared values from multifactor anova designs," *Educational and Psychological Measurement*, vol. 64, pp. 916–924, 2004.
[51] I. H. Witten and E. Frank, *Data Mining - Pratical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufmann Publishers, 2000.

Leandro L. Minku (F.L.Minku 63341860) received the BSc and MSc degrees in Computer Science from the Federal University of Paraná, Brazil, in 2003, and from the Federal University of Pernambuco, Brazil, in 2006, respectively. He is currently working towards the Ph.D. degree in Computer Science at the University of Birmingham, UK. His research interests include on-line learning, concept drift, neural network ensembles and evolutionary computation. Mr. Minku is the recipient of the Overseas Research Students Award (ORSAS) from the British Government (2006) and was the recipient of the Brazilian Council for Scientific and Technological Development (CNPq) scholarships (2006, 2004, 2002 and 2001).

Allan P. White received his Bachelor degree in Psychology in 1973 from the University of Stirling, Scotland and in 1979 obtained an MSc in Computer Science from Newcastle University in the UK. In 1982, he was awarded a PhD in Mathematical Psychology from Durham University, also in the UK. He began his academic career in 1976 with fixed term teaching and research posts at Durham University and in 1979 obtained a tenured post at the University of Birmingham in the UK, where he currently runs the Statistical Advisory Service. He was elected a Fellow of the Royal Statistical Society in 1984. He has written or co-authored about forty papers and articles and his current research interests are machine learning, pattern recognition and forecasting in financial time series.

Xin Yao (M91-SM96-F03) received the BSc degree from the University of Science and Technology of China (USTC), Hefei, Anhui, in 1982, the MSc degree from the North China Institute of Computing Technology, Beijing, in 1985, and the PhD degree from USTC in 1990. He worked as an associate lecturer, lecturer, senior lecturer and associate professor in China and later on in Australia. Currently, he is a professor at the University of Birmingham (UK), a visiting chair professor at the USTC and the director of the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA). He was the editor-in-chief of the IEEE Transactions on Evolutionary Computation (2003-2008), an associate editor or editorial board member of 12 other journals, and the editor of the World Scientific Book Series on Advances in Natural Computation. His major research interests include several topics under machine learning and data mining. Dr. Yao was awarded the President's Award for Outstanding Thesis by the Chinese Academy of Sciences for his PhD work on simulated annealing and evolutionary algorithms in 1989. He also won the 2001 IEEE Donald G. Fink Prize Paper Award for his work on evolutionary artificial neural networks. He is a fellow of the IEEE.