Proceedings of the 25th IASTED International Multi-Conference
ARTIFICIAL INTELLIGENCE AND APPLICATIONS
February 12–14, 2007, Innsbruck, Austria
ISBN Hardcopy: 978-0-88986-629-4 / CD: 978-0-88986-631-7

A FRAMEWORK FOR GENERATING DATA TO SIMULATE CHANGING ENVIRONMENTS

Anand Narasimhamurthy
School of Electronics and Computer Science,
University of Wales, Bangor,
LL57 1UT, United Kingdom
anand@informatics.bangor.ac.uk

Ludmila I. Kuncheva
School of Electronics and Computer Science,
University of Wales, Bangor,
LL57 1UT, United Kingdom
l.i.kuncheva@bangor.ac.uk

ABSTRACT
A fundamental assumption often made in supervised classification is that the problem is static, i.e. the description of the classes does not change with time. However many practical classification tasks involve changing environments. Thus designing and testing classifiers for changing environments are of increasing interest and importance. A number of benchmark data sets are available for static classification tasks. For example, the UCI machine learning repository is extensively used by researchers to compare algorithms across various domains. No such benchmark datasets are available for changing environments. Also, while generating data for static environments is relatively straightforward, this is not so for changing environments. The reason is that an infinite amount of changes can be simulated, and it is difficult to define which ones will be realistic and hence useful. In this paper we propose a general framework for generating data to simulate changing environments. The paper gives illustrations of how the framework encompasses various types of changes observed in real data and also how the two most popular simulation models (STAGGER and moving hyperplane) are represented within.

KEY WORDS
Machine Learning, Changing environments, Concept drift, Population drift, Artificial data, Simulated data.

1 Introduction

Online learning is increasingly occupying centre stage in data mining as more and more applications generate massive streams of data [5, 6, 29]. In these cases it is often impractical to apply conventional classification methodologies since these typically are *batch* algorithms often requiring more than one pass through the entire data. Many online versions of conventional classification methods have been developed [23–25].

A fundamental assumption in supervised classification, including many of the online classification methods, is that the problem is static, i.e. the description of the classes does not change with time. In many practical problems this is most likely to be false. An example is detecting and filtering out spam-email. The descriptions of the two classes "spam" and "non-spam" evolve with time. They are user-specific, and users' preferences change with time. Also spammers are active opponents who devise newer methods to evade detection thereby altering the definition of class "spam" within the current feature space. A change in the environment is referred to as *concept drift* in machine learning. The changes could include minor fluctuations of the underlying probability distributions, steady trends, rapid substitution of one classification task with another and so on [12, 33]. These categories are by no means mutually exclusive and various combinations are possible. Bespoke online classification models [5, 6, 32], as well as classifier ensembles [21, 29] have been proposed for changing environments [22].

In the light of the growing importance of classification in changing environments, it is striking how little attention has been paid to putting together a collection of benchmark data sets or formulating generic procedures for simulating data. As a result researchers use their own simulated datasets or specific real data domains such as finance [8, 12, 32], text categorisation [17–19], spam filtering [4] and web searches [11]. In spite of some critiques [27], the creation of the UCI Machine Learning Repository [3] in 1987 has had a significant positive impact on Machine Learning and Pattern Recognition Communities [1]. A KDD (Knowledge Discovery in Databases) archive was subsequently developed in 1999 [1] containing large data sets of diverse types and domains (`http://kdd.ics.uci.edu`). Data with a variety of known changes are needed for evaluating classification methods in changing environments. In this study we propose a general framework for generating artificial data which simulate changing environments. The data may contain one or more types of drift as desired by the user.

The rest of the paper is organised as follows. In Section 2 we review the terminology used in the literature on changing environments. The proposed framework is described in Section 3. Illustrations of how the proposed framework accommodates the defined types of environmental changes are given in Section 4. An implementation algorithm is given in Section 5. Section 6 concludes the study.

2 Terminology

The most used term in the ML literature on the subject is 'concept drift'. However, there is a variety of other terms, some of them not precisely defined, taken to mean the same thing. This lack of uniformity of terminology prompted us to try to summarise and explain the terms found in the literature before proceeding with the general framework.

The term 'concept' is used to mean slightly different things. It is sometimes used to refer to the class of interest. A concept change in this case would mean a deviation of the description of the class from the original description. Consider for example the STAGGER dataset described by Widmer and Kubat [33] (this is discussed further in section 4.2). In this dataset the feature space of a simple blocks world is described by three features (attributes) namely, size, color and shape. Examples are labelled according to a target concept (this is not given to the classifier explicitly) which changes over time. For example the target concept for the first set of instances is "size = small AND color = red", the target concept for a subsequent set of instances is "color = green OR shape = circular". An analogous 'target concept' or a 'concept of interest' in a real world domain such as information filtering, may be the class 'relevant articles' from a set of documents (e.g. Klinkenberg [14, 18]).

In other articles, the authors use 'concept' to refer to the whole distribution of the problem at time moment t, including all classes. Then concept drift is the substitution of one concept at time t with another at time $t+1$. To avoid confusion, in this study we will try to avoid the term altogether. Another frequently used term is *population drift*, this refers to the changes in the underlying distributions. Kelly et al. [12] use the term *population drift* to mean a changes either in the priors or in the probabilities probabilities of class membership, conditional on feature vectors. They point out that the term *concept drift* has been used in the machine learning literature both for this (population drift) as well as other changes.

While in a natural system we can expect gradual drifts (e.g., seasonal, demographic, habitual, etc.), sometimes the class description may change rapidly due to so called *hidden contexts*. Such contexts may be, for example, illumination in image recognition and accents in speech recognition [34]. The context might instantly become highly relevant. Consider a system trained on images with similar illumination. If images with the same type of content but a different illumination are fed to the system, the class descriptions might change so as to make the system worse than a random guess.

The type of changes can be roughly summarised as follows. They are briefly discussed in subsequent sections. These categories are by no means mutually exclusive and various combinations are possible. For example, recurring trends and population drift are both subsumed by the other categories, namely gradual changes and concept substitution. These are listed separately however, since they are changes which occur in real world applications and also because authors make distinctions.

- Gradual changes (gradual drift, evolutionary changes, concept drift) [2, 10, 13, 16, 30]
- Substitutions (abrupt changes, revolutionary changes, concept substitution, concept shift) [10, 16, 33]
- Recurring trends (recurring contexts) [31, 33]
- Population drift [12]

3 A general framework for simulating changing environments

Every classification problem, however complex it might be, may be described completely by the following. Let $\mathbf{x} \in \Re^n$ be an object in the n-dimensional feature space of the problem and $\Omega = \{\omega_1, \ldots, \omega_c\}$ be the set of class labels. The knowledge of the prior probabilities for the classes, $P(\omega_i)$ and the class-conditional probability density functions (pdf) $p(\mathbf{x}|\omega_i)$, $i = 1, \ldots, c$, determine completely the problem and the optimal (minimum error) classifier for it. Viewed in this probabilistic sense, a classification problem may change due to the changes in the $P(\omega_i)$ and/or $p(\mathbf{x}|\omega_i)$. Posterior probabilities $P(\omega_i|\mathbf{x})$ and the unconditional pdf $p(\mathbf{x})$ may also change but these changes can be re-expressed in terms of $P(\omega_i)$ and $p(\mathbf{x}|\omega_i)$.

Without loss of generality assume that all data live in the n-dimensional real space, $\mathbf{x} \in \Re^n$. Consider a set of K data sources with known distributions. The distribution for source i is characterised by the class-conditional probability density functions $p_i(\mathbf{x}|\omega_j)$, and the prior probabilities $P_i(\omega_j)$, $j = 1, \ldots, c$, for this source, $i = 1, \ldots, K$. At any time we have one or more "active" data sources. Let $v_i(t) \in [0, 1]$ specify the extent of the influence of data source i at time t. We shall treat the influences as mixing proportions and the resultant distribution as a mixture, i.e. $\sum_i v_i(t) = 1$ for any t. At time t the data distribution $D^{(t)}$ is characterised by prior probabilities

$$P(\omega_j, t) = \sum_{i=1}^{K} v_i(t) P_i(\omega_j)$$

and class-conditional pdfs

$$p(\mathbf{x}|\omega_j, t) = \sum_{i=1}^{K} v_i(t) p_i(\mathbf{x}|\omega_j).$$

As the distributions at the sources are fixed, the data distribution at moment t, $D^{(t)}$ is specified through $v_i(t)$. Thus we can equivalently define $D^{(t)}$ as

$$D^{(t)} = \{v_1(t), \ldots, v_K(t)\}.$$

To be able to cater for all possible scenarios of changes we will allow the set of sources to be as large as necessary. For example, if 1000 time instances are needed in a population drift data, the set of sources might need to be of cardinality 1000, one source for each time instance.

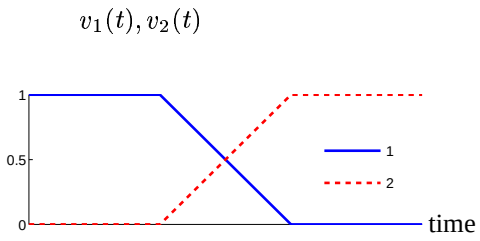

Figure 1. Mixing proportions $v_1(t)$ and $v_2(t)$ for a linear gradual change between two data sources

4 Illustration of the types of changes within the proposed framework

4.1 Gradual drift

Two examples of gradual drift are given below. The first example represents only two sources with gradual changes from one source to the other, as proposed by Widmer and Kubat [33] and Black and Hickey [2]. This is a scenario where one concept fades gradually while the other takes over. The real-world example given by Widmer and Kubat is that of a device which begins to malfunction. At first only small number of data points will come from the stable failure state. Finally the failure will take over completely. Figure 1 shows $v_1(t)$ and $v_2(t)$ as functions of t for this case.

Figure 2 (a) represents the starting data where the distribution is equivalent to that of source 1, i.e., $P(\omega_1) = P(\omega_2) = 0.5$,

$$p(\mathbf{x}|\omega_1) = \begin{cases} \frac{1}{2}, & \text{if } -1 \le x_1 \le 0, -1 \le x_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

$$p(\mathbf{x}|\omega_2) = \begin{cases} \frac{1}{2}, & \text{if } 0 \le x_1 \le 1, -1 \le x_2 \le 1 \\ 0, & \text{otherwise} \end{cases}$$

The end distribution is that of source 2, shown in Figure 2 (d). Subplots (b) and (c) plot the transition of the data distribution for $(v_1(t), v_2(t))$ equal to (0.67,0.33) and (0.33,0.67), respectively.

The second example of gradual change is when there are more sources so the mixture components at any time of the transition are not just the start and the end ones. This example mimics population drift whereby the parameters of the distributions change gradually with time. An example from real-life would be the change of category "high risk" in loan approval. The criteria for a money lender to approve a loan evolve with time to respond to new socio-economic circumstances; the gross annual salaries increase; the market is getting saturated with loan providers. All this will alter the class distributions, so class "high risk" from a few years ago will have a new profile now. To model gradual changes in the parameters of the distributions, we assume that there are as many data sources as there are time instances to be recorded in the transition. At time instance t, $D^{(t)}$ will be a set of binary mixing proportions, with $v_i(t) = 1$ for the data source corresponding to t, and all other $v_j(t) = 0, j \neq i$. To make the model plausible, the consecutive data sources should be closely related. For example, data source for time $t+1$ can be obtained from data source for time t by tweaking a parameter. In this example, the starting and the final distributions are the same as in the first example. However, at each time t, only one source is sampled, so the two classes in the example remain separable at any transition time moment t. Figure 3 shows the scatterplot of the data. Note that this time there are four data sources with their corresponding $v_i(t)$.

This second example of gradual change illustrates a popular artificial data type used by many authors to test their classification methods for changing environments. This data simulation idea is often called *the moving hyperplane* [7, 11, 21].

The proposed framework is meant to be a versatile tool to combine the scenarios of the two examples by letting many data sources to be sampled at a time with mixing proportions evolving with time. It is hoped that in this way the generated data will represent a more realistic model of real-life changing environments.

4.2 Substitutions

This is the type of change most often simulated in order to assess the performance of a classifier in changing environment. The classical example is the STAGGER concept first proposed by Schlimmer and Granger [28]. It has received the status of a benchmark artificial data by being used by many authors since then [4, 7, 20–22, 33].

The feature space is described by three features (attributes): size $\in$ {small (1), medium (2), large (3)}, color $\in$ {red (a), green (b), blue (c)} and shape $\in$ {square (A), circular (B), triangular (C)}. There are three data sources (called 'target concepts')

- Target Concept 1 : size = small AND color = red
- Target Concept 2 : color = green OR shape = circular
- Target Concept 3 : size = medium OR size = large

In their experiment, Widmer and Kubat [33] generate randomly (from a uniform distribution across the feature space) 120 training instances and label each instance according to the current target concept. After processing each instance, the predictive accuracy is tested on an independent test set of 100 instances, also generated randomly. One target concept is active at a time. Concept 1 is active from instance 1 to instance 40, Concept 2 is active from instance 41 to 80 and Concept 3 from instance 81 to 120.

This example is modelled as follows. Consider three data sources each related to one of the target concepts. The distributions at the three data sources are