Pattern Recognition 45 (2012) 521–530

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A unifying view on dataset shift in classification

Jose G. Moreno-Torres [a,*], Troy Raeder [b], Rocío Alaiz-Rodríguez [c], Nitesh V. Chawla [b], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, Universidad de Granada, 18071 Granada, Spain
[b] Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA
[c] Universidad de León, Dpto. de Ingeniería Eléctrica y de Sistemas, Campus de Vegazana, 24071 León, Spain

ARTICLE INFO

Article history:
Received 29 November 2010
Received in revised form
6 June 2011
Accepted 15 June 2011
Available online 18 July 2011

Keywords:
Dataset shift
Data fracture
Changing environments
Differing training and test populations
Covariate shift
Sample selection bias
Non-stationary distributions

ABSTRACT

The field of dataset shift has received a growing amount of interest in the last few years. The fact that most real-world applications have to cope with some form of shift makes its study highly relevant. The literature on the topic is mostly scattered, and different authors use different names to refer to the same concepts, or use the same name for different concepts. With this work, we attempt to present a unifying framework through the review and comparison of some of the most important works in the literature.

1. Introduction

The machine learning community has analyzed data quality in classification problems from different perspectives, including data complexity [29,7], missing values [19,21,39], noise [11,64,58,38], imbalance [52,27,53] and, as is the case with this paper, dataset shift [4,44,14]. Dataset shift occurs when the testing (unseen) data experience a phenomenon that leads to a change in the distribution of a single feature, a combination of features, or the class boundaries. As a result the common assumption that the training and testing data follow the same distributions is often violated in real-world applications and scenarios.

While the research area of dataset shift has received significant attention in recent years (most of the work is published in the last eight years), the field suffers from a lack of standard terminology. Independent authors working under different conditions use different terms, making it difficult to find and compare proposals and studies in the field.

Contributions. The main goal of this work is to provide a unifying framework through the review and analysis of some of the most important publications in the field, comparing the terminology used in each of them and the exact definitions that were given. We present a framework that can be useful in future research and, at the same time, provide researchers unfamiliar with the topic a brief introduction to it. Our goal with this work is to not only unify different methods and terminologies under a taxonomical structure, but also provide a guide to a researcher as well as a practitioner in machine learning and pattern recognition. We use the notation in [44] as the base for the comparisons. We also present a brief summary of solutions proposed in the literature.

The remainder of this paper is organized as follows: Some basic notation is introduced in Section 2. In Section 3, an analysis of the name given to the field of study is presented. Section 4 details the terminology used for the different types of dataset shift that can appear. Section 5 presents examples demonstrating the effect of these shifts on classifier performance. An analysis of some common causes of dataset shift is presented in Section 6. A brief summary of the solutions proposed in the literature is shown in Section 7. Finally, some conclusions are presented in Section 8.

2. Notation

In this work, we focus on the analysis of dataset shift in classification problems. A classification problem is defined by:

- A set of features or *covariates* x.
- A target variable y (the class variable).
- A joint distribution $P(y,x)$.

* Corresponding author.
E-mail addresses: jose.garcia.mt@decsai.ugr.es (J.G. Moreno-Torres), traeder@cse.nd.edu (T. Raeder), rocio.alaiz@unileon.es (R. Alaiz-Rodríguez), nchawla@cse.nd.edu (N.V. Chawla), herrera@decsai.ugr.es (F. Herrera).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.06.019

When analyzing dataset shift, the relationships between the covariates and the class label are particularly relevant. Fawcett and Flach [20] proposed a taxonomy to classify problems according to an intrinsic property of the data generation process: the causal relationship between class label and covariates. This particular characteristic of a problem determines what kinds of shift can affect a given problem, so the rest of the paper is structured regarding the two different kinds of problems generated by this distinction:

- $X \rightarrow Y$ problems, where the class label is causally determined by the values of the covariates. A typical example would be credit card fraud detection, since the behavior of the user, represented in the covariate space X, determines the class label: whether there is fraud or not.
- $Y \rightarrow X$ problems, where the class label causally determines the values of the covariates. Medical diagnosis usually falls in this category, where the disease, which is modeled as the class label Y, determines the symptoms, represented in the machine learning task as covariates X.

The joint distribution $P(y,x)$ can be written as

- $P(y|x)P(x)$ in $X \rightarrow Y$ problems.
- $P(x|y)P(y)$ in $Y \rightarrow X$ problems.

In this prototypical classification problem, the output of the system or learning algorithm takes on N (symbolic) values $y = \{1, \ldots, N\}$ corresponding to N classes. A commonly used loss function for this problem measures the classification error

$$L(y, f(x,\omega)) = \begin{cases} 0 & \text{if } y = f(x,\omega) \\ 1 & \text{if } y \neq f(x,\omega) \end{cases}$$

where ω denotes the set of classifier parameters. Using this loss function, the risk functional

$$R(\omega) = \int L(y, f(x,\omega)) p(x,y)\, dx\, dy$$

quantifies the probability of misclassification. Learning then becomes the problem of estimating the function $f(x,\omega_0)$ (classifier) that minimizes the probability of misclassification using only the training data.

When we use the terms *training* and *test* stages, we refer to the data available to train the classifier and the data present in the environment the classifier will be deployed in, respectively. The data distributions in training and test are denoted as P_{tr} and P_{tst}.

3. Dataset shift

The term "dataset shift" was first used in the book by Quiñonero-Candela et al. [44], the first compilation on the field, where it was defined as "cases where the joint distribution of inputs and outputs differs between training and test stage" [49].

One of the main problems in the field is the lack of visibility most works suffer, since there is not even a standard term to refer to it. So far, each author has chosen a different name to refer to the same basic idea. As an example, the following terms have been used in the literature to refer to dataset shift:

- "Concept shift" or "concept drift" [57,17], where the idea of different data distributions is associated with changes in the class definitions (i.e. the "concept" to be learned).
- "Changes of classification" [55], where it is defined as "In the change mining problem, we have an old classifier, representing some previous knowledge about classification, and a new data set that has a changed class distribution."
- "Changing environments" [4], defined as "The fundamental assumption of supervised learning is that the joint probability distribution $p(x\|d)$ will remain unchanged between training and testing. There are, however, some mismatches that are likely to appear in practice."
- "Contrast mining in classification learning" [60], a slightly different take on the issue: "Given two groups of interest, a user often needs to know the following. Do they represent different concepts? To what degree do they differ? What is the discrepancy and where does it originate from?"
- "Fracture points", defined in [14] as "fracture points in predictive distributions and alteration to the feature space, where a fracture is considered as the points of failure in classifiers' predictions - deviations from the expected or the norm."
- "Fractures between data", used in [40], defined as the case where "we have data from one laboratory (dataset A), and derive a classifier from it that can predict its category accurately. We are then presented with data from a second laboratory (dataset B). This second dataset is not accurately predicted by the classifier we had previously built due to a fracture between the data of both laboratories."

Such inconsistent terminology is a disservice to the field as it makes literature searches difficult and confounds the discussion of this important problem. We recommend the term *dataset shift* for any situation in which training and test data follow distributions that are in some way different. Formally, we define it as

Definition 1. *Dataset shift* appears when training and test joint distributions are different. That is, when $P_{tr}(y,x) \neq P_{tst}(y,x)$.

4. Types of dataset shift

In this section, we present an analysis of the different kinds of shift that can appear in a classification problem. Section 4.1 deals with covariate shift, while Sections 4.2 and 4.3 explain prior probability shift and concept shift, respectively. A graphical example is introduced to illustrate each of these cases. The section is closed with Section 4.4, where other potential types of shifts are explained.

4.1. Covariate shift

The term covariate shift was first defined ten years ago in [47] where it refers to changes in the distribution of the input variables x. Covariate shift is probably the most studied type of shift, but there appears to be some confusion in the literature about the exact definition of the term. There are also some equivalent names, such as "population drift" [31,26]. Some definitions of covariate shift found in the literature are:

- "Case when the population distribution can change over time" (this concept is defined as "population drift" in [31]).
- "Let x be the explanatory variable or the covariate, (…). Let $q_1(x)$ be the density of x for evaluation of the predictive performance, while $q_0(x)$ be the density of x in the observed data. The situation $q_0(x) \neq q_1(x)$ will be called covariate shift in distribution." [47].
- "Change in the data distributions" [26], uses the term 'population drift'.
- "The input distribution $p(x)$ varies but the functional relation $p(y|x)$ remains unchanged" [59].
- "Differing training and test distributions" [8], who define it as follows (the two definitions appear in different places in the same paper):
  - "The training instances are governed by a distribution that is allowed to differ arbitrarily from the test distribution."

  - Training and test distribution may differ arbitrarily, but there is only one unknown target conditional class distribution $p(y|x)$."
- "The conditional probability $p(y|x)$ remains unchanged, but the input distribution $p(x)$ differs from training to future data" [4].
- "The data distribution generating the feature vector x and its related class label y changes as a result of a latent variable t. Thus, we may state that covariate shift has occurred when $P(y|x,t1) \neq P(y|x,t2)$" [14].

The concept of covariate shift is not standardized enough, as can be seen from the differences between the definitions shown above. The definition given by Cieslak and Chawla [14] states that $P(y|x,t1) \neq P(y|x,t2)$, while Yamakazi et al. [59] or Alaiz-Rodríguez et al. [4] state that $p(y|x)$ remains unchanged. Even within the same paper, the two definitions given by Bickel et al. [8] are not equivalent.

In [49], covariate shift is defined as something that occurs "when the data is generated according to a model $P(y|x)P(x)$ and where the distribution $P(x)$ changes between training and test scenarios." This seems to capture the essence of the term as it is most commonly used. Thus, we propose the following as a consistent formal definition.

Definition 2. *Covariate shift* appears only in $X \rightarrow Y$ problems, and is defined as the case where $P_{tr}(y|x) = P_{tst}(y|x)$ and $P_{tr}(x) \neq P_{tst}(x)$.

The analogous issue in $Y \rightarrow X$ problems is prior probability shift, studied in Section 4.2.

Assume we have an $X \rightarrow Y$ problem where there is one covariate x_0 and a target y. The training data distribution $P_{tr}(x_0)$ is composed by the union of two Gaussian distributions with variance 0.5 (one with mean $x_0 = -2$ and the other with mean $x_0 = 2$) and $P_{tr}(y|x_0)$ is defined as

$$P_{tr}(y|x_0) = \frac{1}{1+\exp\left(\frac{-x_0}{0.2}\right)}$$

Consider now that in the test data, $P_{tst}(y|x_0)$ remains unchanged, but the Gaussian distributions that compose $P_{tst}(x_0)$ are now centered in $x_0 = -1$ and $x_0 = 1$, respectively. Fig. 1 depicts this simple example of covariate shift where $P_{tr}(x_0) \neq P_{tst}(x_0)$.

4.2. Prior probability shift

Prior probability shift refers to changes in the distribution of the class variable y. It also appears with different names in the literature, and the definitions have slight differences between them:

- "Change in class distributions" [56], the authors call it "varying class distributions".
- "The class prior probability $p(y)$ varies from training to test, but $p(x|y)$ remains unaltered" [4], denoted as "change in class distribution".
- "Shifting priors occurs when sampling is dependent on the class label and independent of the feature vector x" [14].

Storkey [49] defines prior probability shift as a case where "an assumption is made that a causal model of the form $P(x|y)P(y)$ is valid, (…), the distribution $P(y)$ changes between training and test situations." According to the definitions present in the literature, prior probability shift is the reverse case of covariate shift. More formally, we define it as

Definition 3. *Prior probability shift* appears only in $Y \rightarrow X$ problems, and is defined as the case where $P_{tr}(x|y) = P_{tst}(x|y)$ and $P_{tr}(y) \neq P_{tst}(y)$.

As an example, assume we have a $Y \rightarrow X$ problem with one covariate x_0 and a target y that may take the class values $y=0$ and $y=1$. In the training data, $P_{tr}(y=0) = P_{tr}(y=1) = 0.5$ and $P_{tr}(x_0|y)$ is defined as

$$x_0 = \begin{cases} \mathcal{N}(2,0.5) & \text{when } y=1 \\ \mathcal{N}(-2,0.5) & \text{otherwise} \end{cases}$$

Consider now that in the test data, $P_{tst}(x_0|y=0)$ and $P_{tst}(x_0|y=1)$ remain unchanged, but the class prior probabilities vary, taking the values $P_{tst}(y=1) = 0.70$ and $P_{tst}(y=0) = 0.30$. This example is illustrated in Fig. 2.

Lastly, it is important to mention that prior probabilities are closely related to cost-sensitive learning [54], so techniques from that field are also applicable.

4.3. Concept shift

Concept shift is usually referred to as "concept drift" in the literature; we propose a change in name here for consistency with the above. Even though this type of shift was not mentioned in [44], some other authors have studied it and proposed the following definitions:

- "A changing context can induce changes in the target concepts, producing what is known as concept drift" [57].
- "A user's behaviors and tasks change with time" [34].
- "Changes to the definitions of the classes" [26].
- "$p(y|x)$ changes between the training and test phases" [59], the author used the term "functional relation change"
- "Case where $p(x)$ is not altered but, $p(y|x)$ varies from training to test" [4], denoted as "class definition change".

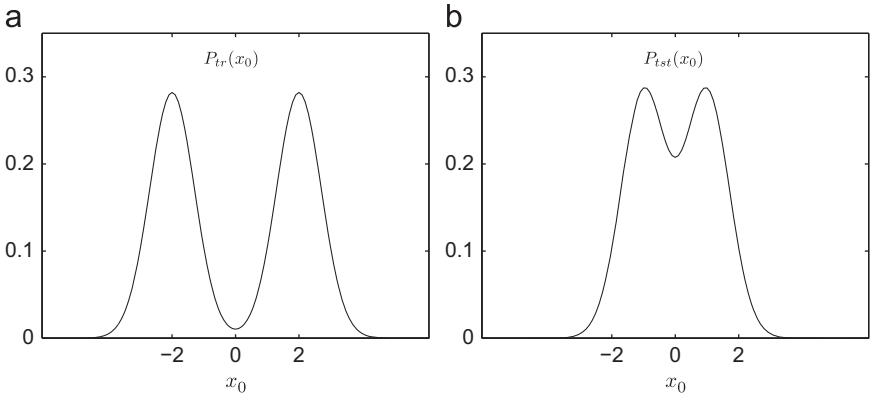

Fig. 1. Covariate shift: $P_{tst}(y|x_0) = P_{tr}(y|x_0)$ and $P_{tr}(x_0) \neq P_{tst}(x_0)$. (a) Training data and (b) test data.

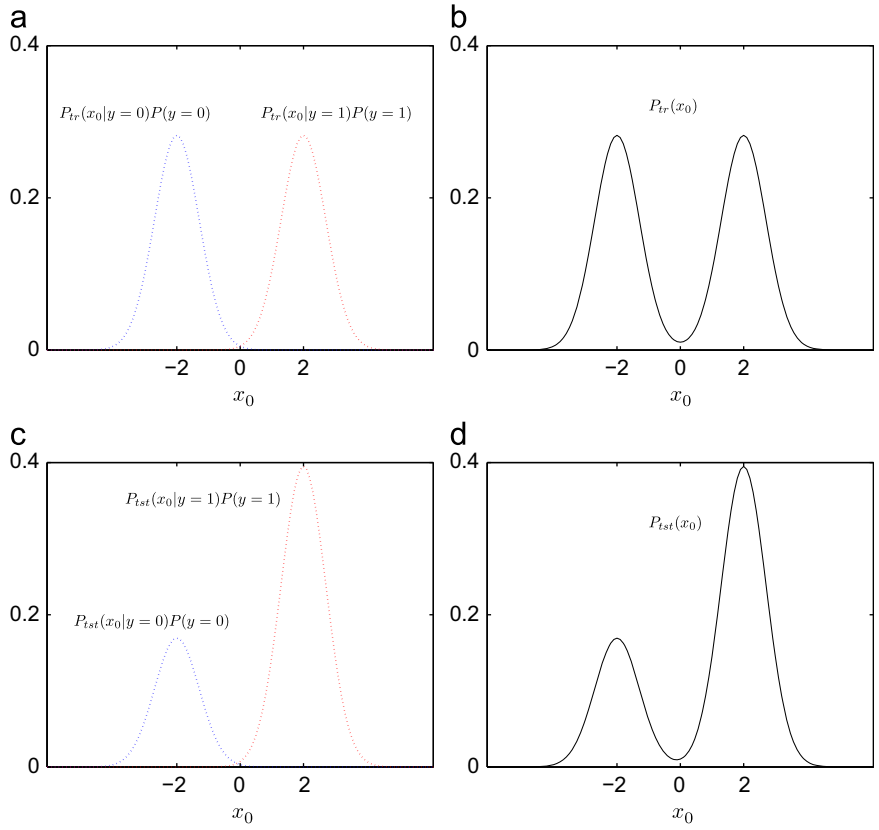

Fig. 2. Prior probability shift. Training dataset with $P_{tr}(y=0)=P_{tr}(y=1)=0.5$. Test dataset with $P_{tr}(y=0)=0.3$ and $P_{tr}(y=1)=0.7$. Class conditional data densities remain constant: $P_{tst}(x_0|y=0)=P_{tr}(x_0|y=0)$ and $P_{tst}(x_0|y=1)=P_{tr}(x_0|y=1)$. (a) Training data, (b) training data density, (c) test data and (d) test data density.

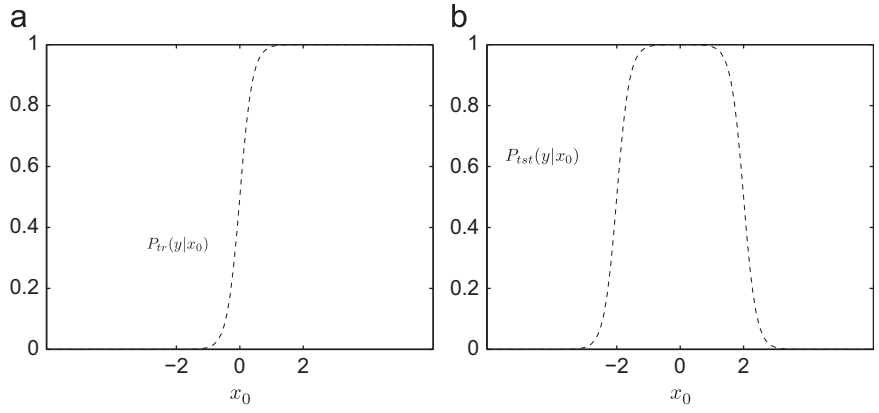

Fig. 3. Example of concept shift: data density remains constant $P_{tr}(x_0)=P_{tst}(x_0)$ and $P_{tr}(y|x_0)\neq P_{tst}(y|x_0)$. (a) Training set and (b) test set.

Concept shift happens when the relationship between the input and class variables changes, which presents the hardest challenge among the different types of dataset shift that has been tackled so far. Formally, we define it as

Definition 4. *Concept shift* is defined as

- $P_{tr}(y|x)\neq P_{tst}(y|x)$ and $P_{tr}(x)=P_{tst}(x)$ in $X\rightarrow Y$ problems.
- $P_{tr}(x|y)\neq P_{tst}(x|y)$ and $P_{tr}(y)=P_{tst}(y)$ in $Y\rightarrow X$ problems.

As an example of concept shift, consider the training dataset with the distribution presented for the covariate shift problem. If a concept shift takes place, the test set data distribution $P_{tst}(x_0)$ remains constant, but $P_{tst}(y|x_0)$ is redefined, for instance, as

$$P_{tst}(y|x_0)=\frac{1}{\left(1+\exp\left(\frac{-2+x_0}{0.2}\right)\right)\left(1+\exp\left(\frac{-2-x_0}{0.2}\right)\right)}$$

Fig. 3 shows the $P_{tr}(y|x_0)$ and $P_{tst}(y|x_0)$ for a concept shift problem.

4.4. Other types of dataset shift

Even though the shifts presented above are the most commonly present in real-world classification tasks, there are others

that could in theory also happen, included here for completeness:

- $P_{tr}(y|x) \neq P_{tst}(y|x)$ and $P_{tr}(x) \neq P_{tst}(x)$ in $X \rightarrow Y$ problems.
- $P_{tr}(x|y) \neq P_{tst}(x|y)$ and $P_{tr}(y) \neq P_{tst}(y)$ in $Y \rightarrow X$ problems.

There are two main reasons these shifts are usually not considered in the literature: they appear more rarely than the others and, most importantly, they are so hard that we currently consider them impossible to solve.

5. Examples of the relevance of dataset shift

The examples presented in Sections 4.1 and 4.2 were designed to showcase as clearly as possible what covariate and prior probability shift mean. However, they do not show why its study is important: the negative effect dataset shift often has on classifier performance.

This section presents new examples for both covariate shift and prior probability shift, where the said shifts actually produce a change in the Bayes error boundary.

Fig. 4 depicts a case of covariate shift where the shift produces a change in the Bayes error boundary resulting in a drop in the classifier performance. In this example, assume we have an $X \rightarrow Y$ problem where there is one covariate x_0 and a target class label y that takes the values $y=0$ and $y=1$. In the training data, $P_{tr}(x_0)$ is composed by the union of two Gaussian distributions, $\mathcal{N}(-1.5,0.5)$ and $\mathcal{N}(1.5,0.5)$, that are the data distributions of each class, respectively. In the test data, $P_{tst}(y|x_0)$ remains unchanged, but the Gaussian distributions that compose $P_{tst}(x_0)$ now have means -1.5 and 0.5, respectively. Fig. 4(d) shows the difference between the optimal decision boundary (continuous line) in the test set and that one estimated from the training dataset (dashed line).

Fig. 5, on the other hand, shows a case of prior probability shift. For this example, assume we have a $Y \rightarrow X$ problem with a covariate x_0 and a target y. In the training data, $P_{tr}(y=0) = P_{tr}(y=1) = 0.5$ and $P_{tr}(x_0|y)$ is defined as

$$x_0 = \begin{cases} \mathcal{N}(1.5,0.5) & \text{when } y=1 \\ \mathcal{N}(-1.5,0.5) & \text{otherwise} \end{cases}$$

In the test data, $P_{tst}(x_0|y)$ remains unchanged, but the prior probabilities change to $P_{tst}(y=1)=0.8$ and $P_{tst}(y=0)=0.2$. Fig. 5 illustrates this problem and Fig. 5(d) highlights the difference between the optimal decision boundary (continuous line) and the boundary estimated in the training stage. If the class prior probabilities differ from the ones assumed during learning, the classifier performance will be suboptimal.

6. Causes of dataset shift

In this section we comment on some of the most common causes of dataset shift. These concepts have created confusion at times, so it is important to remark that these terms are factors that can lead to the appearance of some of the shifts explained in Section 4, but they do not constitute dataset shift themselves.

There are several possible causes for dataset shift, out of which this section mentions the two we deem most important: Sample selection bias and non-stationary environments. In the first one, the discrepancy in distribution is due to the fact that the training examples have been obtained through a biased method, and thus do not represent reliably the operating environment where the classifier is to be deployed (which, in machine learning terms, would constitute the test set). This case is studied in Section 6.1, and is the one most commonly analyzed in the literature.

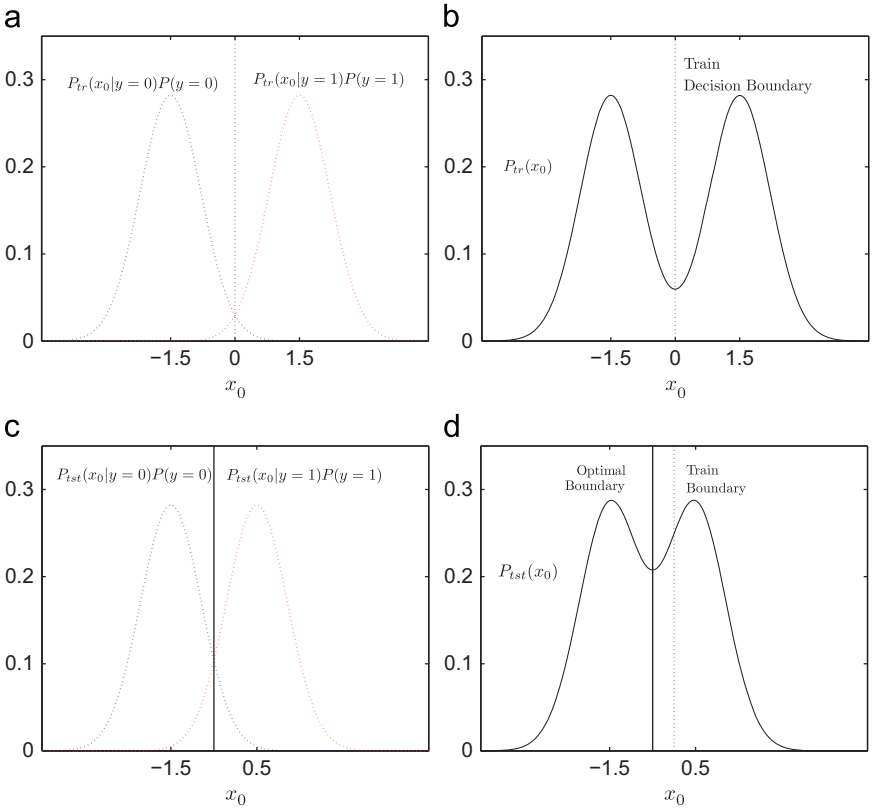

Fig. 4. Example of covariate shift with an influence on the Bayes error boundary. The vertical dotted line represents the boundary learned by the classifier using the training set. The vertical continuous line represents the optimal boundary for the test set. (a) Training set, (b) training data density, (c) test set and (d) test data density.

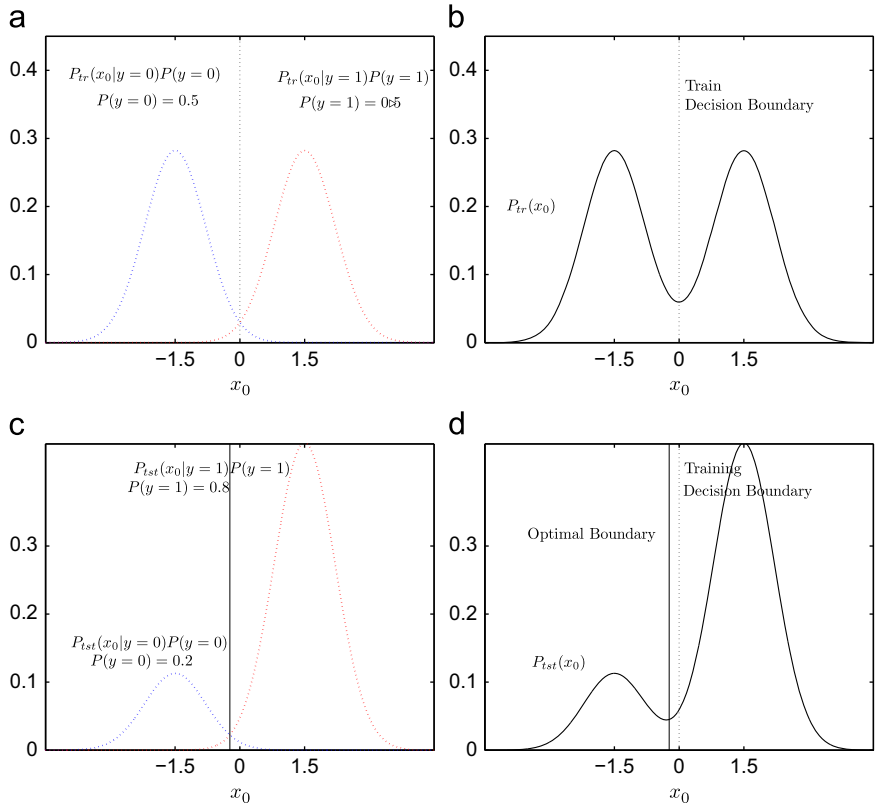

Fig. 5. Example of prior probability shift with an influence on the Bayes error boundary. The vertical dotted line represents the boundary learned by the classifier using the training set. The vertical continuous line represents the optimal boundary for the test set. (a) Training set, (b) training data density, (c) test set and (d) test data density.

A typical example of this case would be the analysis of a process where, due to cost concerns, one of the classes is sampled at a lower rate than it actually appears.

The second cause appears when the training environment is different from the test one, whether it is due to a temporal or a spatial change. It commonly appears, among others, in adversarial classification problems; and it is analyzed in Section 6.3.

6.1. Sample selection bias

The term *sample selection bias* refers to a systematic flaw in the process of data collection or labeling which causes training examples to be selected *non-uniformly* from the population to be modeled. In social science research, for example, there will be subsets of the general population (students at the researcher's University or previous study participants) which are easier to survey than others. These "easy" populations may be over-represented in the training sample, whereas "difficult" populations (i.e. prisoners) may be under-represented or completely excluded.

One can imagine any number of permutations of this general problem. If data are collected from remote sensors, for example, the different sensors may malfunction at different rates or collect data at different rates, meaning that certain portions of the observation area are over-represented.

The problem of operating under sample selection bias has received substantially more attention in other domains than it has in the machine learning community. In the credit scoring literature it goes by the name of *reject inference*, because potential credit applicants who are *rejected* under the previous model are not available to train future models [15,25].

The term has been used as a synonym of covariate shift [30] (which is not correct, as was stated above), but also on its own as a related problem to dataset shift. In that line, Storkey [49] proposes the following formal definition:

Definition 5. *Sample selection bias*, in general, causes the data in the training set to follow $P_{tr}=P(s=1|x,y)$, while the data in the test set follows $P_{tst}=P(y,x)$. Depending on the type of problem, we have:

- $P_{tr}=P(s=1|y,x)P(y|x)P(x)$ and $P_{tst}=P(y|x)P(x)$ in $X\rightarrow Y$ problems,
- $P_{tr}=P(s=1|y,x)P(x|y)P(y)$ and $P_{tst}=P(x|y)P(y)$ in $Y\rightarrow X$ problems,

where s is a binary selection variable that decides whether a datum is included in the training sample process ($s=1$) or rejected from it ($s=0$).

In [37,61,14], three different types of sample selection bias were analyzed:

Definition 6. *Missing completely at random (MCAR)* occurs when the sampling method is completely independent of x and y, so that $P(s=1|y,x)=P(s=1)$. This kind of bias does not produce any dataset shift.

Definition 7. *Missing at random (MAR)* occurs when s depends on x but conditional on x is independent of y; so that $P(s=1|y,x)=P(s=1|x)$. This kind of bias can potentially produce covariate shift.

To illustrate more clearly the relationship between MAR bias and covariate shift, note that one can "correct" for covariate shift when estimating model performance by using *importance-weighted cross-validation* [51]. That is to say, an *unbiased estimate* of the classification loss on a set of feature vectors x_i and their associated classes y_i can be obtained by weighting the loss associated with each x_i by $P_{tst}(x_i)/P_{tr}(x_i)$. More formally, if the k-fold cross-validation

estimate of misclassification cost is given by

$$\frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{F}_j|}\sum_{i=1}^{|\mathcal{F}_j|}\ell(x_i,y_i,\hat{y}_i) \quad (1)$$

where $\ell(\,)$ represents the classification loss incurred by the classification estimate $\hat{y}_i$[1] on the instance with covariates x_i and class y_i, then a "nearly unbiased" estimate of the classification loss under covariate shift can be computed as

$$\frac{1}{k}\sum_{j=1}^{k}\frac{1}{|\mathcal{F}_j|}\sum_{i=1}^{|\mathcal{F}_j|}\frac{P_{tst}(x_i)}{P_{tr}(x_i)}\ell(x_i,y_i,\hat{y}_i) \quad (2)$$

Here the term "nearly unbiased" means that the estimate becomes unbiased as the sample size $n\to\infty$. In the case of leave-one-out cross-validation, IWCV provides an unbiased estimate of the classification loss for a dataset with $n-1$ samples [51].

Under MAR bias, we have that $P_{tr}(x_i)=P(s=1|x_i)P_{tst}(x_i)$, meaning that $P_{tst}(x_i)/P_{tr}(x_i)=P(s=1|x_i)^{-1}$. Thus, "correcting" for MAR bias under simple loss functions amounts to estimating $P(s=1|x_i)$. This estimation can be accomplished in practice by building a classifier to predict $F:\mathbf{x}\to s$, that is, building a classifier with s as the class label. Such a construction is often feasible in practical applications. In credit scoring, for example, we only know the class label y (default) of applicants for whom $s=1$ (meaning credit was approved). However, creditors retain the application information for all applicants even those for whom $s=0$ (credit is denied) [5,61].

Effective correction of MAR bias, then, reduces to the problem of producing a *well-calibrated classifier* which predicts $P(s=1|x_i)$ as accurately as possible. In general this is not trivial, as many algorithms (such as Naive Bayes and Boosting) have been shown to produce probabilities that are skewed toward 0 or 1 [41,63].

Definition 8. *Missing not at random (MNAR)* occurs when there is no independence assumption between x, y and s. This kind of bias can introduce one or more of covariate shift, prior probability shift and concept shift.

Under MNAR bias, the selection mechanism may depend on the class attribute as well as the observed features. The most famous method for correcting MNAR bias comes from Heckman [28] who shows how to estimate a linear model over both observed and unobserved data when the dependent variable is known only for the observed data. Specifically, assume we have linear models for both the class variable y and the selection variable s of the form:

$$y_i=\beta_1 x_{1i}+u_{1i}$$

$$s_i=\beta_2 x_{2i}+u_{2i}$$

$$u1,u2\sim N(\mathbf{0},\sigma_{u1}^2,\rho) \quad (3)$$

Here the two β_j are 1-by-k_j model parameter vectors and the two x_{ji} are k_j-by-1 feature vectors for individual instances i. The vector x_{1i} contains the features upon which the class value depends, and x_{2i} contains the features on which the selection process depends. Thus, in Heckman's model, the class and selection variables are linear in some feature space with potentially correlated Gaussian noise.

Heckman proves that with these assumptions, an unbiased model y^* for the entire dataset can be built with the following procedure:

1. Estimate the parameters of the model s_i by some method such as ordinary least squares.
2. Set $\lambda_i=\phi(x_{2i}\beta_2)/\Phi(x_{2i}\beta_2)$.
3. Estimate the parameters of a new linear model y^* which includes λ as an independent variable.

Here ϕ and Φ are the standard normal PDF and CDF, respectively. Zadrozny and Elkan [62] generalize this procedure for arbitrary classification tasks by building one classifier to predict the selection label s and incorporating that classifier's predictions into a second classifier for predicting the class label y. While this approach has no theoretical guarantees, it was shown to be effective in a real-world application.

For completeness sake, we have defined a fourth option to be considered:

Definition 9. *Missing at random-class* (*MARC*) occurs when s depends on y but conditional on y is independent of x; so that $P(s=1|y,x)=P(s=1|y)$. This kind of bias can potentially produce prior probability shift.

Sufficient and necessary conditions for sample selection bias: Quiñonero-Candela et al. [44] give a set of conditions that the densities P_{tr} and P_{tst} need to satisfy in order for the classification problem to be modeled as a sample selection bias problem, meaning that its training and test densities can be expressed as in Definition 5. These conditions can be stated as follows:

1. *Support condition* $P_{tr}(x)>0\to P_{tst}(x)>0$.
2. *Selection condition* $\sup_x(P_{tr}(x,y)/P_{tst}(x,y)<\infty)$.

The support condition simply states that any feature vector x that can be drawn from the training distribution can also be drawn from the test distribution. The selection condition is slightly stronger, requiring that any pair (x,y) of a feature vector and class label that can be drawn from the $P_{tr}(x,y)$ can also be drawn from $P_{tst}(x,y)$. Fig. 6 explains this graphically. The red histogram shows a potential test density, the black histogram is a training density that may have been generated by sample selection bias (its density is nonzero everywhere the test density is nonzero) and the blue histogram shows a density that must be modeled by some other form of dataset shift.

This observation exposes a key difference between sample selection bias and covariate shift. Even in the case (MAR) where $p(s=1)$ depends only on the feature vector x, the framework of sample selection bias imposes a *stricter* criterion on the relationship between P_{tr} and P_{tst} than covariate shift. That is to say, there are some instances of covariate shift that cannot arise from MAR bias, but every instance of MAR bias can be modeled as covariate shift (Fig. 6(a)). As such, any technique that is developed to correct for covariate shift should also be able to correct for MAR bias, but the reverse is not true.

6.2. Challenges in correcting sample selection bias

We have seen that many established techniques to compensate for sample selection bias depend critically on the estimation of the selection variable s. In the case of IWCV, we need a well-calibrated estimate of $P(s=1|x)$ while the Zadrozny and Heckman techniques require a monotonic score. In either case if the chosen model is a poor fit, the correction procedure will be ineffective and may degrade rather than improve model performance [48].

If the feature sets x_1 and x_2 are identical (i.e. the same features are used to estimate both s and y), then the additional variable λ may end up highly correlated with the "uncorrected" estimate y_1. In this case, the Heckman procedure has little power to correct for sample selection bias. Little and Rubin [36] state that the Heckman procedure requires "significant" predictive variables in x_2 that are

[1] It is worth noting that the "classification estimate" may be real-valued, such as an estimate of $p(1|x)$.

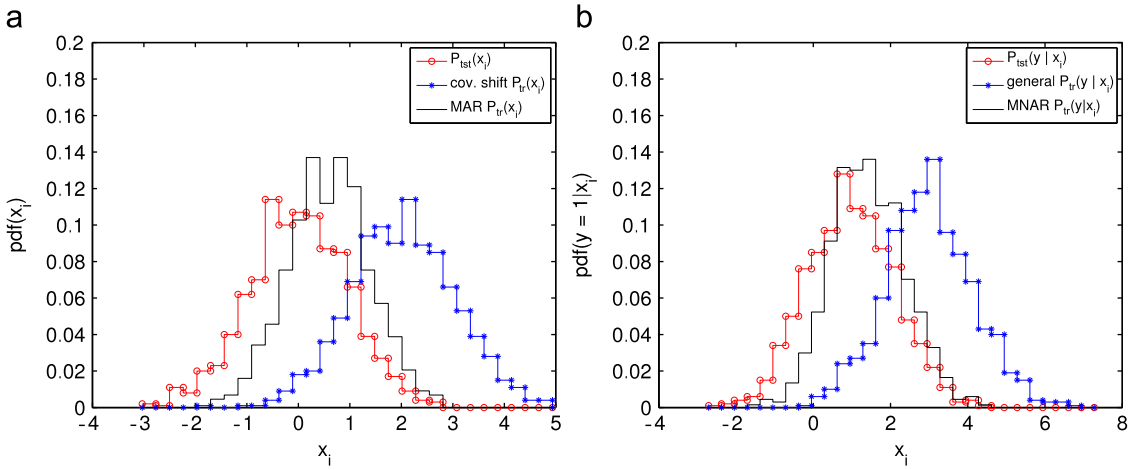

Fig. 6. Sufficient and necessary conditions for sample selection bias. The red curve shows a test pdf and the black and blue curves show potential training pdfs. The black density may be modeled as sample selection bias. The blue curve violates the (a) support condition and (b) selection condition. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

not in x_1 in order to be effective in many cases [43]. A broader survey of critiques to the Heckman correction can be found in [43].

When attempting to correct MAR bias with techniques such as importance-weighted cross-validation, one may run into trouble if $P(s=1|x_i)=0$. This situation, often referred to as *censorship*, arises when a deterministic procedure (such as a credit model) determines the value of s. Censorship may be addressed by modeling the problem as MNAR regardless of any explicit dependency on the class label y [12].

6.3. Non-stationary environments

In real-world applications, it is often the case that the data is not (time- or space-) stationary. Depending on the type of problem, non-stationary environments can introduce different kinds of shift:

- In $X \rightarrow Y$ problems, a non-stationary environment could create changes in either $P(x)$ or $P(y|x)$, generating covariate shift or concept shift, respectively.
- In $Y \rightarrow X$ problems, it could generate prior probability shift with a change in $P(y)$ or concept shift with a change in $P(x|y)$.

One of the most relevant non-stationary scenarios involves adversarial classification problems, such as spam filtering and network intrusion detection. This type of problem is receiving an increasing amount of attention in the machine learning field [16,6,10,35], and usually copes with non-stationary environments due to the existence of an adversary that tries to work around the existing classifier's learned concepts. In terms of the machine learning task, this adversary warps the test set so that it becomes different from the training set, thus introducing any possible kind of dataset shift.

There are also other applications where non-stationariness appears. They include remote sensing applications, where a dataset collected in a given season for an area with different terrains is employed to train the classifier but, when that classifier is deployed, mismatches may appear due to seasonal changes or because the new region has a different terrain distribution [3]; direct mail marketing, where the proportion of target customers or customer profiles may vary from one city to the next; and biometric authentication, among others.

7. Proposals in the literature for the analysis of dataset shift

In this section we give a brief overview of the different proposals that have appeared in the literature to work under the different types of dataset shift.

Covariate shift has been extensively studied in the literature, and a number of proposals to work under it have been published. Some of the most important ones include weighting the log-likelihood function [47], importance-weighted cross-validation [51], asymptotic Bayesian generalization error [59], discriminative learning [9], kernel mean matching [23], or adversarial search [22].

Prior probability shift has also been studied deeply, with a multitude of proposals appearing in the literature. There are two main strategies when designing classifiers for expected prior probability shift conditions:

- *Adaptive approaches*: These proposals train a classifier over the available data and then the adapt some of its parameters according to the (usually unlabeled) test data. This adaptation may be done either by the end user [33,31] or automatically [46,3].
- *Robust approaches*: Base the choice of classifier on some measure that is ideally transparent to changes in class distribution. The best known example would be ROC curve analysis [1,42] (which has generated some controversy, see [56,20]), but there are others too [18,2]. The automatic choice of classifier parameters [32] can also be considered a robust approach.

Other significant proposals in the literature have focused on determining the existence and/or shape of dataset shift between two datasets. Wang et al. [55] present the idea of correspondence tracing. They propose an algorithm for the discovering of changes of classification characteristics, which is based on the comparison between two rule-based classifiers, one built from each dataset. Yang et al. [60] present the idea of conceptual equivalence as a method for contrast mining, which consists of the discovery of discrepancies between datasets. Chawla and coworkers [14,45] developed a statistical framework to analyze changes in data distribution resulting in fractures between the data.

Lastly, there are some approaches that try and modify the data to repair dataset shift. Among them, Klinkenberg [32] proposed an example selection/weighting approach and Moreno-Torres et al. [40] applied a GP-based feature extraction technique to repair fractures between data originated in different biological laboratories by finding a transformation over the data from one of the laboratories.

8. Concluding remarks

In many practical applications of machine learning, the data available for model-building (training data) are not strictly representative of the data on which the classifier will ultimately be deployed (test data). This problem, which we call *dataset shift* in accordance with [44] generalizes a wide variety of researches that are scattered throughout the machine learning literature. The purpose of this paper is to survey and unify this research in order to better inform future endeavors in the field.

Researchers studying the general problem of dataset shift, or specific instances of this problem, have coined a number of different names for it. These include *concept shift* [57], *concept drift* [57], *covariate shift* [47], *data fracture* [14,40], *reject inference* [24,15], and *imprecise class distributions* [2], among others. Worse still, researchers have sometimes used different terms to refer to the same problem, or given different definitions to the same term. To clear up this confusion and to make future research easier, we have carefully studied the terminology used in the literature and proposed a common convention which attempts to capture the essence of the terms as they are most commonly used. Specifically, we propose:

- *Covariate shift* if $P_{tst}(x) \neq P_{tr}(x)$ but $P_{tst}(y|x) = P_{tr}(y|x)$, in accordance with [47].
- *Prior probability shift* if $P_{tst}(y) \neq P_{tr}(y)$ but $P_{tst}(y|x) = P_{tr}(y|x)$.
- *Concept shift* if $P_{tst}(x) = P_{tr}(x)$ but $P_{tst}(y|x) \neq P_{tr}(y|x)$ (in $X \rightarrow Y$ problems) or $P_{tst}(x|y) \neq P_{tr}(x|y)$ (in $Y \rightarrow X$ problems).
- *Dataset shift* if $P_{tst}(x,y) \neq P_{tr}(x,y)$ but none of the above hold.

Next, we survey common causes of dataset shift. *Sample selection bias* [12,28,61] occurs when the training sample is selected non-uniformly at random from the test population. Depending on the selection criteria and the type of classification problem, selection bias may produce covariate shift, prior probability shift, or general dataset shift. In *adversarial* environments [10,16,35] such as spam detection and fraud detection, *adversaries* continually adapt the test data to the output of the classification algorithm. The adversaries try to produce data (with some constraints) which the learner will misclassify as often as possible. This tends to produce general dataset shift as the adversary may alter the test distribution arbitrarily. In *non-stationary* environments, the dataset shift arises from a significant physical or temporal difference between training and test data sources. If a model trained on one continent is applied on another, for example, arbitrary changes in data distribution may result.

Finally, we have briefly surveyed some proposals in the literature for learning under dataset shift, either *detecting* that a shift has occurred or *adapting* to the shift once it does occur. We plan to expand on this in much greater detail in future work.

Acknowledgments

Jose García Moreno-Torres is currently supported by an FPU Grant from the Ministerio de Educación y Ciencia of the Spanish Government. This work was supported in part by the Spanish Government's KEEL project (TIN2008-06681-C06-01). This work was also supported in part by the National Science Foundation (NSF) Grant ECCS-0926170. Lastly, the work was also partially supported by the Spanish projects DPI2009-08424 and TEC2008-01348/TEC.

References

[1] N.M. Adams, D.J. Hand, Comparing classifiers when the misallocation costs are uncertain, Pattern Recognition 32 (7) (1999) 1139–1147.

[2] R. Alaiz-Rodríguez, A. Guerrero-Curieses, J. Cid-Sueiro, Minimax regret classifier for imprecise class distributions, Journal of Machine Learning Research 8 (2007) 103–130.

[3] R. Alaiz-Rodríguez, A. Guerrero-Curieses, J. Cid-Sueiro, Classification under changes in class and within-class distributions, in: Proceedings of the 10th International Work-Conference on Artificial Neural Networks, IWANN '09, Springer-Verlag, Berlin, Heidelberg, 2009, pp. 122–130.

[4] R. Alaiz-Rodríguez, N. Japkowicz, Assessing the impact of changing environments on classifier performance, in: Proceedings of the Canadian Society for Computational Studies of Intelligence, 21st Conference on Advances in Artificial Intelligence, Canadian AI '08, Springer-Verlag, Berlin, Heidelberg, 2008, pp. 13–24.

[5] J. Banasik, J. Crook, L. Thomas, Sample selection bias in credit scoring models, Journal of the Operational Research Society 54 (8) (2003) 822–832.

[6] M. Barreno, B. Nelson, A.D. Joseph, J.D. Tygar, The security of machine learning, Machine Learning (2010) 121–148.

[7] M. Basu, T.K. Ho, Data Complexity in Pattern Recognition, Springer-Verlag Inc., New York, Secaucus, NJ, USA, 2006.

[8] S. Bickel, M. Brückner, T. Scheffer, Discriminative learning for differing training and test distributions, in: Proceedings of the 24th International Conference on Machine Learning, ICML 2007, ACM, New York, NY, USA, 2007, pp. 81–88.

[9] S. Bickel, M. Brückner, T. Scheffer, Discriminative learning under covariate shift, Journal of Machine Learning Research 10 (2009) 2137–2155.

[10] B. Biggio, G. Fumera, F. Roli, Multiple classifier systems for robust classifier design in adversarial environments, International Journal of Machine Learning and Cybernetics 1 (2010) 27–41.

[11] C.E. Brodley, P. Uiversity, M.A. Friedl, B. Uiversity, B.P. Edu, Identifying mislabeled training data, Journal of Artificial Intelligence Research 11 (1999) 131–167.

[12] N. Chawla, G. Karakoulas, Learning from labeled and unlabeled data: an empirical study across techniques and domains, Journal of Artificial Intelligence Research 23 (1) (2005) 331–366.

[14] D.A. Cieslak, N.V. Chawla, A framework for monitoring classifiers' performance: when and why failure occurs? Knowledge and Information Systems 18 (1) (2009) 83–108.

[15] J. Crook, J. Banasik, Does reject inference really improve the performance of application scoring models? Journal of Banking & Finance 28 (4) (2004) 857–874.

[16] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma, Adversarial classification, in: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '04, ACM, New York, NY, USA, 2004, pp. 99–108.

[17] T.G. Dietterich, G. Widmer, M. Kubat, Special issue on context sensitivity and concept drift, Machine Learning 32 (2) (1998).

[18] C. Drummond, R.C. Holte, Explicitly representing expected cost: an alternative to ROC representation, in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 198–207.

[19] A. Farhangfar, L. Kurgan, J. Dy, Impact of imputation of missing values on classification error for discrete data, Pattern Recognition 41 (12) (2008) 3692–3705.

[20] T. Fawcett, P.A. Flach, A response to Webb and Ting's 'on the application of ROC analysis to predict classification performance under varying class distributions', Machine Learning 58 (1) (2005) 33–38.

[21] M. Ghannad-Rezaie, H. Soltanian-Zadeh, H. Ying, M. Dong, Selection–fusion approach for classification of datasets with missing values, Pattern Recognition 43 (6) (2010) 2340–2350.

[22] A. Globerson, C.H. Teo, A. Smola, S. Roweis, An adversarial view of covariate shift and a minimax approach, in: J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence (Eds.), Dataset Shift in Machine Learning, The MIT Press, 2009, pp. 179–198.

[23] A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, B. Schölkopf, Covariate shift by kernel mean matching, in: J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence (Eds.), Dataset Shift in Machine Learning, The MIT Press, 2009, pp. 131–160.

[24] D. Hand, Reject inference in credit operations, in: Credit risk modeling: design and application, 1998, pp. 181–190.

[25] D. Hand, W. Henley, Statistical classification methods in consumer credit scoring: a review, Journal of the Royal Statistical Society: Series A 160 (3) (1997) 523–541.

[26] D.J. Hand, Rejoinder: classifier technology and the illusion of progress, Statistical Science 21 (1) (2006) 30–34.

[27] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284.

[28] J. Heckman, Sample selection bias as a specification error, Econometrica: Journal of the Econometric Society (1979) 153–161.

[29] T.K. Ho, M. Basu, Complexity measures of supervised classification problems, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (3) (2002) 289–300.

[30] J. Huang, A.J. Smola, A. Gretton, K.M. Borgwardt, B. Schölkopf, Correcting sample selection bias by unlabeled data, Advances in Neural Information Processing Systems 19 (2007) 601–608.

[31] M.G. Kelly, D.J. Hand, N.M. Adams, The impact of changing populations on classifier performance, in: Proceedings of the Fifth ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining, KDD 99, 1999, pp. 367–371.
[32] R. Klinkenberg, Learning drifting concepts: example selection vs. example weighting, Intelligent Data Analysis 8 (3) (2004) 281–300.
[33] M. Kubat, R.C. Holte, S. Matwin, Machine learning for the detection of oil spills in satellite radar images, Machine Learning 30 (2–3) (1998) 195–215.
[34] T. Lane, C.E. Brodley, Approaches to online learning and concept drift for user identification in computer security, in: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, KDD, AAAI Press, 1998, pp. 259–263.
[35] P. Laskov, R. Lippmann, Machine learning in adversarial environments, Machine Learning 81 (2010) 115–119.
[36] R. Little, D. Rubin, Statistical Analysis with Missing Data, 1987.
[37] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, Probability and Statistics, second ed., Wiley, New Jersey, 2002.
[38] Z.-Y. Liu, H. Qiao, Multiple ellipses detection in noisy environments: a hierarchical approach, Pattern Recognition 42 (11) (2009) 2421–2433.
[39] J. Luengo, S. García, F. Herrera, A study on the use of imputation methods for experimentation with Radial Basis Function Network classifiers handling missing attribute values: the good synergy between rbfns and eventcovering method, Neural Networks 23 (3) (2010) 406–418.
[40] J.G. Moreno-Torres, X. Llorà, D.E. Goldberg, R. Bhargava, Repairing fractures between data using genetic programming-based feature extraction: a case study in cancer diagnosis, Information Sciences, in press, doi:10.1016/j.ins.2010.09.018.
[41] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: Proceedings of the ICML, ACM, 2005, pp. 625–632.
[42] F. Provost, T. Fawcett, Robust classification for imprecise environments, Machine Learning 42 (3) (2001) 203–231.
[43] P. Puhani, The Heckman correction for sample selection and its critique, Journal of Economic Surveys 14 (1) (2000) 53–68.
[44] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence, Dataset Shift in Machine Learning, The MIT Press, 2009.
[45] T. Raeder, N.V. Chawla, Model monitor: evaluating, comparing, and monitoring models, Journal of Machine Learning Research 10 (2009) 1387–1390.
[46] M. Saerens, P. Latinne, C. Decaestecker, Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure, Neural Computation 14 (1) (2002) 21–41.
[47] H. Shimodaira, Improving predictive inference under covariate shift by weighting the log-likelihood function, Journal of Statistical Planning and Inference 90 (2) (2000) 227–244.
[48] R. Stolzenberg, D. Relles, Tools for intuition about sample selection bias and its correction, American Sociological Review 62 (3) (1997) 494–507.
[49] A. Storkey, When training and test sets are different: characterizing learning transfer, in: J. Quiñonero Candela, M. Sugiyama, A. Schwaighofer, N.D. Lawrence (Eds.), Dataset Shift in Machine Learning, The MIT Press, 2009, pp. 3–28.
[51] M. Sugiyama, M. Krauledat, K.-R. Müller, Covariate shift adaptation by importance weighted cross validation, Journal of Machine Learning Research 8 (2007) 985–1005.
[52] Y. Sun, M.S. Kamel, A.K. Wong, Y. Wang, Cost-sensitive boosting for classification of imbalanced data, Pattern Recognition 40 (12) (2007) 3358–3378.
[53] Y.M. Sun, A.K.C. Wong, M.S. Kamel, Classification of imbalanced data: a review, International Journal of Pattern Recognition and Artificial Intelligence 23 (4) (2009) 687–719.
[54] K.M. Ting, A study on the effect of class distribution using cost-sensitive learning, in: Fifth International Conference on Discovery Science, DS 2002, 2002, pp. 98–112.
[55] K. Wang, S. Zhou, C.A. Fu, J.X. Yu, F. Jeffrey, X. Yu, Mining changes of classification by correspondence tracing, in: Proceedings of the 2003 SIAM International Conference on Data Mining (SDM 2003), 2003, pp. 95–106.
[56] G.I. Webb, K.M. Ting, On the application of ROC analysis to predict classification performance under varying class distributions, Machine Learning 58 (1) (2005) 25–32.
[57] G. Widmer, M. Kubat, Learning in the presence of concept drift and hidden contexts, Machine Learning 23 (1996) 69–101.
[58] X. Wu, X. Zhu, Mining with noise knowledge: error aware data mining, IEEE Transactions on SMC, Part A 28 (4) (2008) 917–932.
[59] K. Yamazaki, M. Kawanabe, S. Watanabe, M. Sugiyama, K.-R. Müller, Asymptotic Bayesian generalization error when training and test distributions are different, in: Proceedings of the 24th International Conference on Machine Learning, ICML '07, ACM, New York, NY, USA, 2007, pp. 1079–1086.
[60] Y. Yang, X. Wu, X. Zhu, Conceptual equivalence for contrast mining in classification learning, Data & Knowledge Engineering 67 (3) (2008) 413–429.
[61] B. Zadrozny, Learning and evaluating classifiers under sample selection bias, in: Proceedings of the 21st International Conference on Machine Learning, ICML '04, ACM, New York, NY, USA, 2004, p. 114.
[62] B. Zadrozny, C. Elkan, Learning and making decisions when costs and probabilities are both unknown, in: Proceedings of the KDD, ACM, 2001, pp. 204–213.
[63] B. Zadrozny, C. Elkan, Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers, in: Proceedings of the ICML, 2001, pp. 609–616.
[64] X. Zhu, X. Wu, Class noise vs attribute noise: a quantitative study, Artificial Intelligence Review 22 (3) (2004) 177–210.

Jose G. Moreno-Torres received the M.Sc. degree in Computer Science in 2008 from the University of Granada, Spain. After spending a year as a fellow of an international "la Caixa" scholarship, during which he did research at the IlliGAL laboratory under the supervision of Prof. David E. Goldberg, he is currently a Ph.D. candidate under the supervision of Prof. Francisco Herrera, working with the Soft Computing and Intelligent Information Systems Group in the Department of Computer Science and Artificial Intelligence at the University of Granada. His current research interests include dataset shift, imbalanced classification, bibliometrics and multi-instance learning.

Troy Raeder is a Ph.D. student in Computer Science at the University of Notre Dame in South Bend, IN, USA. His research interests include scenario analysis in machine learning, evaluation methodologies in machine learning, and robust models for changing data distributions. He received B.S. and M.S. degrees in Computer Science from Notre Dame in 2005 and 2009, respectively.

Rocío Alaiz-Rodríguez received the B.S. degree in Electrical Engineering from the University of Valladolid, Spain, in 1999 and the Ph.D. degree from Carlos III University of Madrid, Spain. She is currently an Associate Professor at the Department of Electrical and Systems Engineering, University of Leon, Spain. Her research interests include learning theory, statistical pattern recognition, neural networks and their applications to image processing and quality assessment (in particular, food and frozen–thawed animal semen).

Nitesh V. Chawla is an Associate Professor in the Department of Computer Science and Engineering at the University of Notre Dame. He directs the Data Inference Analysis and Learning Lab (DIAL) and is also the co-director of the Interdisciplinary Center of the Network Science and Applications (iCenSA) at Notre Dame. His research is supported with research grants from organizations such as the National Science Foundation, the National Institute of Justice, the Army Research Labs, and Industry Sponsors. His research group has received numerous honors, including best papers, outstanding dissertation, and a variety of fellowships. He has also been noted for his teaching accomplishments, receiving the National Academy of Engineers CASEE New Faculty Fellowship, and the Outstanding Undergraduate Teacher Award in 2008 and 2011. He is an Associated Editor for IEEE Transactions of Systems, Man and Cybernetics Part B and Pattern Recognition Letters. More information is available at http://www.nd.edu/~nchawla.

Francisco Herrera received his M.Sc. degree in Mathematics in 1988 and Ph.D. degree in Mathematics in 1991, both from the University of Granada, Spain. He is currently a Professor in the Department of Computer Science and Artificial Intelligence at the University of Granada. He has had more than 200 papers published in international journals. He is coauthor of the book "Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases" (World Scientific, 2001). He currently acts as Editor in Chief of the international journal "Progress in Artificial Intelligence" (Springer) and serves as Area Editor of the Journal Soft Computing (area of evolutionary and bioinspired algorithms) and International Journal of Computational Intelligence Systems (area of information systems). He acts as Associated Editor of the journals: IEEE Transactions on Fuzzy Systems, Information Sciences, Advances in Fuzzy Systems, and International Journal of Applied Metaheuristics Computing; and he serves as a member of several journal editorial boards, among others: Fuzzy Sets and Systems, Applied Intelligence, Knowledge and Information Systems, Information Fusion, Evolutionary Intelligence, International Journal of Hybrid Intelligent Systems, Memetic Computation, Swarm and Evolutionary Computation. He received the following honors and awards: ECCAI Fellow 2009, 2010 Spanish National Award on Computer Science ARITMEL to the "Spanish Engineer on Computer Science", and International Cajastur "Mamdani" Prize for Soft Computing (Fourth Edition, 2010). His current research interests include computing with words and decision-making, data mining, bibliometrics, data preparation, instance selection, fuzzy rule-based systems, genetic fuzzy systems, knowledge extraction based on evolutionary algorithms, memetic algorithms and genetic algorithms.