Automatic correction of SVM for drifted data classification

Alexandra Degeest*,**, Benoît Frénay*, Michel Verleysen*

*Machine Learning Group, ICTEAM, Université Catholique de Louvain,
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium
**ISIB, Haute-Ecole Paul-Henri Spaak
Rue Royale 150, 1000 Bruxelles, Belgium
alexandra.degeest@uclouvain.be, degeest@isib.be

Abstract. Concept drift is an important feature of real-world data streams that can make usual machine learning techniques rapidly become unsuitable. This paper addresses the problem of sudden concept drift in classification problems for which standard techniques may fail. To this end, support vector machines (SVMs) are automatically corrected to cope with a new suddenly drifted dataset. Results on real-world datasets with several types of sudden drift indicate that the method is able to correct the SVM in order to better classify the new data after the concept drift, using a correction based on the difference between the initial dataset and the new drifted dataset, even when the new dataset is small.

1 Introduction

Concept drift in real-world data streams is a standard concern. Concept drift happens when the distribution underlying the data changes over time. It can occur in several ways: sudden drift (abrupt modification of the data statistical distribution), gradual drift (period of time when two different distributions are active), incremental drift (progressive modification of the data statistical distribution) or reoccurring contexts (periodic reappearance of different statistical distributions) (Zliobaite, 2010). This paper deals with sudden drift. The reasons why sudden concept drift happens in datasets are numerous and diverse. As an example, in industrial applications, data can suddenly drift because of engine maintenance, because the tool has been relocated or because the new data have been collected by another person. The occurrence of concept drift in data streams often disturbs usual machine learning techniques and, therefore, brings the necessity to develop new learning techniques or to adapt existing ones.

Support vector machines (SVMs) are a powerful tool, commonly used for classification of high-dimensional data. However, SVMs may quickly become useless when data are subject to concept drift. The objective of this paper is to propose a method to automatically correct the SVM model facing sudden drift in a classification problem. The goal is twofold. First, the goal is to allow us using the first dataset (before the drift) for model learning, what is particularly useful when the second dataset (after the drift) is small. The second goal is to correct the decision hyperplane to better classify new datasets, even when their statistical distribution (slightly) differs from the distribution of the first dataset. Two assumptions have been made in this work. First, the labels of the drifted data are unknown. Second, the drifted dataset is small

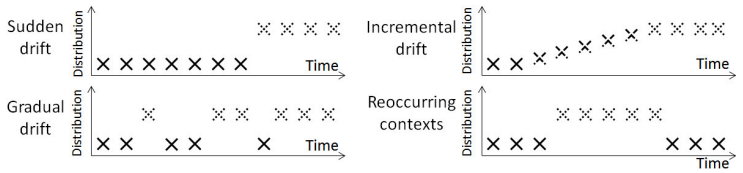

FIG. 1 – *Structural types of concept drift, inspired by Zliobaite (2010).*

compared to the first dataset. Section 2 is a brief literature review about the different types of concept drift. The problem and the proposed method are respectively described in Sections 3 and 4. Section 5 presents the experimental results while Section 6 concludes the work.

2 Concept Drift

Concept drift happens when the data distribution changes over time. This phenomenon frequently happens in real-world data streams. To cope with concept drift, new machine learning approaches must be developed or common techniques must be improved. Several thorough overviews about concept drift have been written (Zliobaite, 2010; Tsymbal, 2004). According to Zliobaite (2010), concept drift techniques should be more adapted to the problem itself because too many studies are looking for a universal answer to drifts in general, what is not optimal enough. Indeed there exist several types of concept drift (see Figure 1), such as sudden drift, gradual drift, incremental drift and reoccurring contexts (Zliobaite, 2010). Each type of drift is associated to a specific class of real-world problems and needs a specific approach.

This paper copes with sudden drift problems; these are common in industry and medicine, for instance. Sudden drift may occur because of engine maintenance, machine relocation or undesired modification in environmental conditions. For example if a physician records an ECG from a first patient on Monday and from a second patient on Tuesday, environmental conditions may have changed (temperature, probe locations...). The two datasets properties (distribution...) will differ, possibly resulting in a drift. However it would be really interesting to be able to use the dataset from the first patient together with the dataset from the second patient, knowing how difficult it can be to obtain a sufficiently large training dataset.

3 Problem Statement

This paper addresses the problem of sudden concept drift in classification problems. The main question is "How to keep and correct a model M1, learned from database D1, in order to use it to evaluate a drifted concept on another database D2?" Two main hypotheses are made in this work. The first is that the labels of the drifted data D2 are unknown; this corresponds to realistic situations. For the first dataset D1, before the drift, a specialist has been available to classify the results during the training and validation periods. However, for the second dataset D2, after the drift, e.g. after a tool relocation, no specialist is available anymore and it would be too expensive to call an expert to give the labels every time there is a slight drift on the statistical distribution. The second hypothesis is that the drifted dataset D2 is much smaller

than the undrifted dataset D1. This also corresponds to a common situation: after the drift, the model must still perform well, even if the new dataset D2 is only composed by a few instances. Therefore, it is interesting to find a methodology that works on a small unlabelled dataset.

SVMs are already used for concept drift detection (Campigotto and al., 2010; Dries and al., 2009; Klinkenberg and al., 2000) and for incremental learning (Rüping, 1999). The closest work to this paper, by Yang and al. (2007), also uses SVM to face sudden drift. The fundamental difference between Yang and al. (2007) and this work is that Yang and al. (2007) needs labeled data in the drifted dataset to adapt the classifier. In addition, it should be noticed that Yang and al. (2007) needs a preprocessing consisting of selecting some instances in the dataset, what might be risky if the dataset is small.

The originality of the method described in this paper is that it adapts directly the model hyperplane to unlabeled drifted data, based on the drifted and undrifted data distributions. There is no need for a hazardous selection of data from D2, and there is no need for labels on these drifted data. This has the advantage of offering a simple and fast correction to the model, once the data have drifted. It is fast because the model does not need to be trained again; in particular the meta-parameters of the model do not need to be selected and validated again, after the concept drift, what can be time-consuming.

4 Support Vector Machine Correction with Drifted Data

This section introduces the main contribution of this paper. A method to correct SVM models facing sudden drift is described. To correct SVM models, the method uses a correction based on the difference between the initial dataset and the new drifted dataset. Let D1 represent the undrifted dataset with labeled instances and D2 the drifted dataset without labels. The SVM model M1 learns only from D1. The meta-parameters of the SVM are selected by cross-validation. Then the classification of M1 over D1 is evaluated. Without modifying M1, the model is evaluated on the drifted dataset D2, possibly resulting in a larger misclassification rate, due to the sudden drift. The principle of the proposed corrective algorithm is then to use the decision values of M1 over D1, called M1(D1) and of M1 over D2, called M1(D2). The decision values represent the distances between the separating hyperplane and each instance of the dataset. Each instance (x, y) of datasets D1 consists of a feature vector $x \in R^n$ and a class label y. Let ω be the weight vector and b be the threshold. The distance between the hyperplane and each instance of the dataset is then $|\omega x + b|/||\omega||$ where $\omega x + b$ is the hyperplane equation, leading to the decision function $f(x) = sign(\omega x + b)$.

Before drift, the optimal hyperplane is the one with the maximal margin of separation between the classes of the undrifted dataset D1. The decision values M1(D2) represent the distance between the same hyperplane, computed on D1, and each instance of the dataset D2. Once the decision values have been computed on D1 and D2, their probability density functions are estimated on both sets. Figure 2 shows an example, with a straight line for M1(D1) and a dashed line for M1(D2). Probability density functions of decision values are used because they are an easy way to characterize in a one-dimensional plot the differences between the probability density functions of M1(D1) and M1(D2).

When the probability density functions have been estimated, the distance between them can be measured. To this end, the integral of the squared difference between the two density functions is calculated. The next step of the procedure is to modify the model in order to better

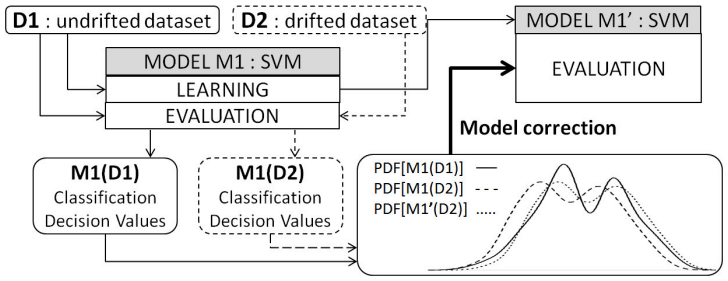

FIG. 2 – *Methodology to correct an SVM model facing sudden drift; see text for details.*

fit the data in D2. For this purpose, the decision values M1(D2) are shifted by α (becoming M1′(D2)) to minimize the distance between M1′(D2) and M1(D1). Shifting the decision values is equivalent to shifting the SVM hyperplane perpendicularly to itself, or, equivalently, to shift the D2 data in the same direction. The optimal value of α is found by gradient descent on the distance between the two probability density functions. This correction factor is then used to correct the SVM hyperplane in order to obtain the corrected model M1′. Applying the new model M1′ on data D2 is expected to result in significantly improved performances compared to using model M1 on D2. The new decision function of M1′ is $f'(x) = sign(\omega x + b + \alpha)$. Figure 2 describes the whole method. The plain arrows represent the use of undrifted data D1 and the dashed arrows represent the use of drifted data D2.

5 Experiments

The experiments in this section show that moving the SVM hyperplane, based on a corrective factor α, without any additional learning, can achieve fast and good results when facing sudden drift, even on a small dataset D2 and when no labeled data is available in this dataset.

For these experiments, two real databases, *Birds* (Jacques and al., 2010) and *Crabs* (Campbell and al., 1974) have been used. The first database *Birds* consists of birds from the same species with different geographical origins. Five morphological variables have been measured on 206 birds of each species. From these variables, the model learns to classify the birds into two groups: males and females. The second database *Crabs* has 100 rows and 6 columns, describing 5 morphological measurements on crabs of both sexes. The model also learns to classify the birds into two groups: males and females. On these databases, different amplitudes of concept drift have been artificially added, in order to validate the proposed methodology.

***Birds* Database.** For the first experiment, the size of D2 has been limited to 80 instances while D1 has kept its 206 instances. The 80 instances of D2 have been selected randomly among the two classes (males and females) and the initial proportion of the classes have been respected. On these instances, seven different amplitudes of concept drift have been added, from a very small drift to an important modification of the data. For each feature, the (fixed) drift corresponds to 9, 18, 24, 26, 30, 36 and 47% of the feature standard deviation. Figure 3

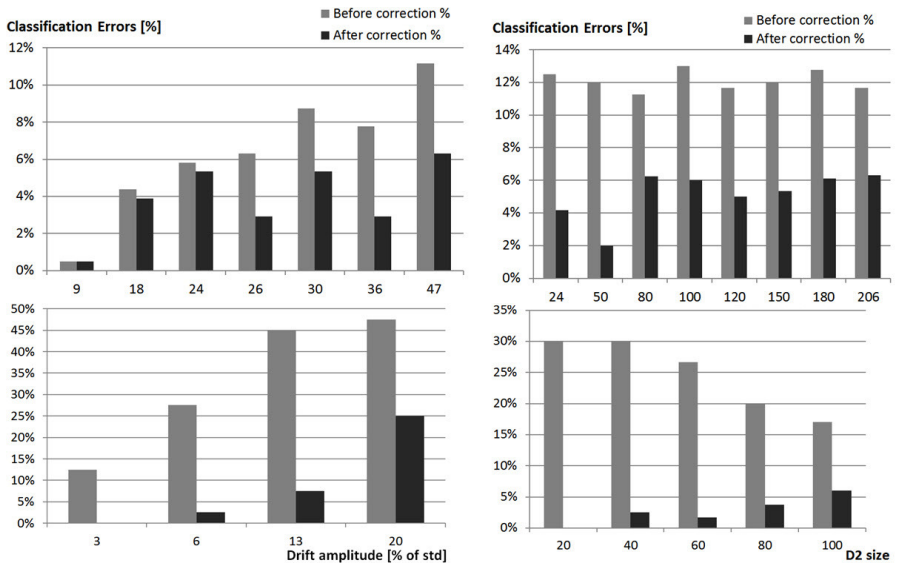

FIG. 3 – *Left: Classification errors for several amplitudes of drift on Birds (top) and Crabs (bottom) databases, before correction (gray) and after correction (black) of the model M1 over the drifted dataset D2. Right: Classification errors for several sizes of D2.*

shows the improvement in classification performances after the SVM correction has been applied. For the second experiment, a fixed amplitude of drift has been set to a reasonably realistic 20% of the features standard deviation in D2. Eight different sizes of dataset D2 have been used: 24, 50, 80, 100, 120, 150, 180, and 206. Figure 3 shows that the classification error percentage drops after the correction for all dataset sizes.

***Crabs* Database.** Similar experiments have been realized on *Crabs* database. For the first experiment, four different amplitudes of concept drift have been added to D2: 3, 6, 13 and 20% of the feature standard deviation. The size of D2 has been set to 40 instances while D1 has kept its 100 instances. The 40 instances of D2 have been selected randomly among the two classes and the initial proportion of the classes have been respected. Figure 3 shows that a good percentage improvement after the correction has been achieved for all drift amplitudes. For the second experiment, a fixed amplitude of drift has been kept: 15% of the standard deviation of the features. Several sizes of D2 dataset have been used: 20, 40, 60, 80 and 100. Figure 3 shows that the classification error percentage drops after the correction for each dataset size.

6 Conclusions

This paper proposes a method to automatically correct the support vector machine algorithm facing sudden drift in a classification problem. To classify correctly the new unlabeled drifted data, the proposed method allows to use the model learned over the initial data (be-

fore drift) and to directly correct its decision hyperplane using the statistical distribution of the drifted instances. The experiments have shown that moving the SVM hyperplane, based on a corrective factor, without any additional learning nor meta-parameter validation, can achieve simple and good results when facing sudden drift.

Résumé

Le concept drift est une caractéristique importante des flux de données réelles qui peut vite rendre les techniques classiques de machine learning inadaptées. Cet article traite du concept drift soudain dans les problèmes de classification où les techniques classiques peuvent échouer. A cette fin, les SVMs sont automatiquement corrigés face au nouvel ensemble de données soudainement driftées. Les résultats sur des bases de données réelles avec différents types de drift soudain montrent que la méthode est capable de corriger le SVM afin de mieux classifier les nouvelles données après le drift, en utilisant une correction basée sur la différence entre l'ensemble initial de données et le nouvel ensemble drifté, même si celui-ci est petit.