Multi One-Class Incremental SVM for Document Stream Digitization

Anh Khoi Ngo Ho, Véronique Eglin, Nicolas Ragot, Jean-Yves Ramel

Multi One-Class Incremental SVM for Document Stream Digitization

Anh Khoi NGO HO, Veronique EGLIN
Laboratoire LIRIS, INSA de Lyon
Lyon, France

Nicolas RAGOT, Jean-Yves RAMEL
Laboratoire d'Informatique (LI EA 6300),
Tours, France.

***Abstract*— Inside the DIGIDOC project (ANR-10-CORD-0020) - CONTenus et INTeractions (CONTINT), our approach was applied to several scenarios of classification of image streams which can correspond to real cases in digitization projects. Most of the time, the processing of documents is considered as a well-defined task: the classes (also called concepts) are defined and known before the processing starts. But in real industrial workflows of document processes, it may frequently happen that the concepts can change during the time. In a context of document stream processing, the information and content included in the digitized pages can evolve over the time as well as the judgment of the user on what he wants to do with the resulting classification. The goal of this application is to create a module of learning, for a steam-based document images classification (especially dedicated to a digitization process with a huge volume of data), that adapts different situations for intelligent scanning tasks: adding, extending, contracting, splitting, or merging the classes in on an online mode of streaming data processing.**

Keywords: machine learning, intelligent scanner, dynamic classification, streaming data, one-class, SVM.

I. INTRODUCTION OF MULTI ONE-CLASS INCREMENTAL SVM

Quite often, the structure of the classifiers and learning algorithms are not flexible enough to allow the introduction of classes that are not expected. In such process, all data are not necessary available from the beginning, because of the context of the streaming data process. Incremental learning has been mainly considered to handle such situations where the number of concepts can vary from the beginning of the process (where we just have partial information and few examples) to the end (where all examples have been processed) ([1], [3]). Another area of dynamic machine learning deals with the resolution of concept drift or non-stationary environment problems ([4], [5]). It represents the ability of the system to have self-adapted to changes that can appear in the description of concepts over the time and to make forget the "old parts" of the concept. More recently, the combination between two problems has been studied in [2] and in [6] but not yet applied in document area, precisely for document classification and recognition tasks. Especially, one can notice that it does not exist any dynamic system able to consider in one unique solution the questions about adding, extending, contracting, merging or splitting concepts.

The main difficulties when a classifier has to deal with evolving concepts come from two points. The first one is the architecture of the system, often static and rigid, mostly not configured to integrate changes in the data classes, as it appears when classes must be added, discarded, split, or merged. The second point is the interdependency of the parameters, especially in discriminant approaches, since they determine decision boundaries relatively to several classes. Consequently, when a concept is changing, all parameters have to be re-adapted to the context of all other data which needs at least many examples (and mostly a full retraining). This kind of mechanism does not adapt to situations dealing with the data stream. The strong relationship between data that is exploited in this family of discriminant classifiers is for sure one of the strength of this kind of methods. Going further in simplifying the interdependencies leads naturally to one-class classifiers that only use positive data from one class to build their model. These approaches are less studied since they are known to be less accurate than discriminative ones. However, they have a huge potential such as: insensibility to imbalanced data, rejection potential, automatic detection to new concepts or ambiguities (and thus potential merging of concepts), multiclass labeling, individual feature space representation, etc.

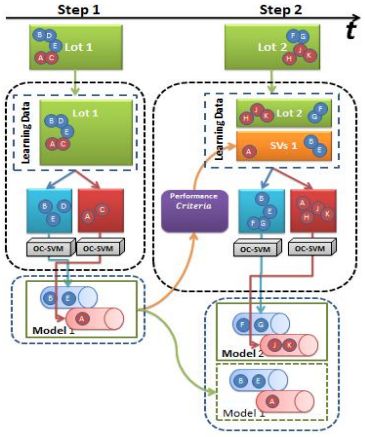

Figure 1. Schema of mOC-iSVM.EP

Our proposed approach, called mOC-iSVM (multi One-Class incremental SVM) is the solution that lies on a set of

one-class SVMs, each one modeling a concept. The originality of our proposal comes from the use of the former knowledge kept in the SVM models (represented by all selected support vectors) and its combination with the new data coming incrementally from the stream. The algorithmic details of the incremental one class classifiers can be found in [7]. For our purpose, the proposed classification model is available in different versions lying on different ways of considering the pre-existent support vectors in each class models: the mOC-iSVM.AP [8] model based on the selection of ancient support vectors according to their «age»; and the mOC-iSVM.EP (Figure 1) model based on the selection of ancient support vectors according to their efficiency. Each one has advantages and limitations depending on the characteristic of the environments (stationary or non-stationary) of the problem.

II. EXPERIMENTS

The results (Table 1) prove that our approach is very interesting in different environments (stationary and non-stationary) compared with others approaches in the domain. We applied exactly the protocol proposed in [2] with *SEA Dataset*. This dataset is designed especially for testing concept drift (non-stationary), included both ***extending, contracting concepts***. And we reuse the experiment protocol in [3] with the *Optical Recognition of Handwritten Digits Dataset* (*ORHD Dataset*) which is designed for testing the ability of ***adding concepts*** during the ***incremental*** process.

Approaches	*ORHD Dataset*	*SEA Dataset*	*Approaches*	*ORHD Dataset*	*SEA Dataset*
mOC-iSVM.AP	98.6 %	96.0 %	**SEA [4]**	82.3 %	95.7 %
mOC-iSVM.EP	98.6 %	97.2 %	**DWM [5]**	82.3 %	96.6 %
Learn++ [1]	75.9 %	--	**Adaboost**	--	93.2 %
Learn++.NC[3]	93.4 %	--	**Bagging**	93.3 %	--
Learn++.NSE [2]	--	96.8%			

Table 1. Results of our approaches

We also test the results of our standard version on both online and batch mode (lot of 100 and 10 data per step) to show the ability to adapt different process of scanning in the reality with the ORHD Dataset [3]. We can make the observation that the performance of the classifier is quite independent of the choice of the lot size (although there is a slight profit in favor of larger lots, which is not surprising)- **Figure 2**.

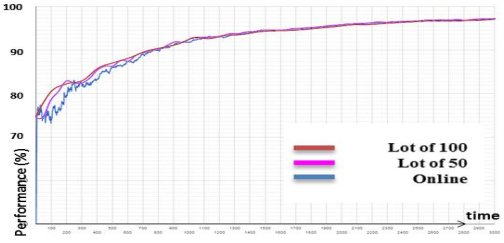

Figure 2. Online and batch mode (*ORHD Dataset*)

We applied this solution to implement a simulation application on a real online scenario to an intelligent scanner on the *DIGIDOC* dataset with the objective to classify documents according to their content, Figure 3. The dataset is composed of almost six different categories of documents (handwriting, printing, map, musical scores) that are supposed to be digitized in a random order and recognize on the streaming line. This application is also able to adapt the user requests (adding, extending, contracting, merging or splitting concepts) via an interactive interface for scanners. The demonstration (video) of a simple scenario of adding and extending classes is available at {https://goo.gl/K5MY3O}.

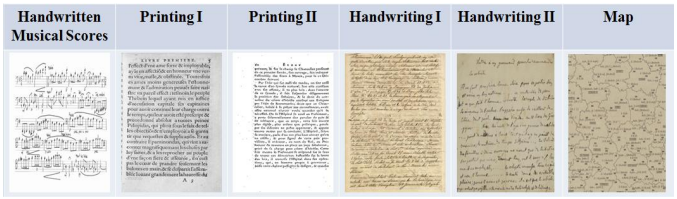

Figure 3. Sample of incoming document in the steam.

III. CONCLUSION

In this contribution, we present a new approach of document classification based on a multi one-class incremental SVM dedicated to document stream recognition. The system has been designed for offering new solutions for digitization with a huge volume data: diagnostic of quality, content recognition… We show also the efficiency of the multi one-class classifier on both stationary and non-stationary environment. The simulation of an intelligent scanner is also proposed. Future works will consist in reinforcing the user interactions and the integration of the dynamic classifier in a real scanner for resolving industrial or patrimonial challenges.

REFERENCES

[1] R.Polikar, L.Udpa, S.Udpa, V.Honavar, "An incremental learning algorithm for supervised neural networks". IEEE Trans. on SMC (C), Special Issue on Knowledge Management, (2001).

[2] R.Elwell, R.Polikar: "Incremental Learning of Concept Drift in Nonstationary Environments". IEEE Trans on Neural Networks (2011).

[3] M.Muhlbaier, A.Topalis, R.Polikar, "Learn++.NC: Combining Ensemble of Classifiers Combined with Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes," IEEE Trans on Neural Networks, vol. 20, no. 1, pp. 152 – 168, (2009).

[4] W.N.Street, Y.S.Kim, "A streaming ensemble algorithm (SEA) for large-scale classication", ACM SIGKDD, (2001).

[5] J.Kolter, M.Maloof ; "Dynamic Weighted Majority (DWM): an ensemble method for drifting concepts" ; JMLR8, 2755–2790 (2007).

[6] M.Bouillon, É.Anquetil, A.Almaksour, 'Decremental Learning of Evolving Fuzzy Inference Systems Using a Sliding Window'. ICMLA'12, Florida, USA (2013).

[7] Anh Khoi Ngo Ho, Nicolas Ragot, Jean-Yves Ramel, Veronique Eglin, Nicolas Sidere, "Document classification in a non-stationary environment: a one class SVM approach", ICDAR'13, USA (2013).

[8] Anh Khoi Ngo Ho, Nicolas Ragot, Jean-Yves Ramel, Veronique Eglin, 'Multi One-Class Incremental SVM For Both Stationary And Non-Stationary Environment', CAp 2014, France (2014).