Full Incremental Learning for Along Classification of Textual Images

Vincent Poulain d'Andecy
L3i, University of La Rochelle
and ITESOFT Group – Yooz
La Rochelle, France
email:vincent.poulaindandecy@yooz.fr

Aurélie Joseph and Saddok Kebairi
Research and Technologies Department
ITESOFT Group – Yooz
Aimargues, France
email:{aurelie.joseph, saddok.kebairi}@yooz.fr

Abstract— **Incremental classification is still a challenge with an important industrial impact by allowing a class training process simplification. Recently, works on Incremental Growing Neural Gas (IGNG) have demonstrated the ability of this technology to cope with this challenge for Optical Character Recognition(OCR)-based image classification. Previous proposals focused on the classifier itself but did not deal with descriptors which were not in the scope of these studies taking an a priori fixed descriptors set. This assumption is not applicable in real-life when the environment is progressive and the incremental system does not know a priori the image content to learn. In this paper we proposed an enhancement of an incremental system based on an IGNG extension (A2ING) with a combination of graphical, re-using the Blurred Shape Model (BSM), and a novel strategy based on incremental textual descriptors. Performance achievement shows a better precision with an acceptable recall than predefined descriptors. The benefit is to not require a prior descriptors selection.**

Keywords-incremental classification; text-based vector; shape-based vector; BSM; A2ING; Document Image Processing.

I. Introduction

Even if more and more documents are managed by electronic exchanges, the paper document is still used and need an image capture for automatic processing. Moreover, the increasing use of mobile devices generates nowadays a large volume of images which can be digitized papers (receipts, etc.) or natural scene image with text (a board, an advertisement, etc.) all so-called documents. It is an industrial challenge to classify all these images for indexing, archiving or business processing. In this paper, we are interested in supervised image classification containing textual information.

Currently, we use an OCR-based system with a supervised classifier. Each class is known and for the training, we have representative images preprocessed by the OCR. A word-class pair weight is calculated to extract specific words featuring the document classes. Thus, learning algorithms calculate the proximity between images and predict which class the document belongs to. We can use different standard algorithms: Support Vector Machine (SVM), Naïves Bayes, K-Nearest Neighbour (k-NN).

To set-up these systems, we need an a priori groundtruth with labelled document by class. These methods are usually efficient but we face to limitations and new needs:

- adding easily a new class in the system,
- discovering new data along the process,
- reducing the number of sample,
- processing big data.

Incremental classification approaches [9] can theoretically manage these issues. Similarly to supervised approaches, incremental classification systems require a feature vector to model the problem. For instance, an image can be represented by its pixels. For a text-based document, the vector could be a bag-of-words. As we will see for this last case, an issue lies in the selection of these words. The number of descriptors is problematic as well. How many have to be selected to represent our vector, knowing that the vector length cannot be growing?

Proposals on incremental classification are numerous [6][10][14]. We chose to use the Active Incremental Growing Neural Gas (A2ING) algorithm which is one of most recent proposals. Our contribution is not yet on the A2ING itself but we propose a novel method based on the A2ING to classify incrementally textual document without a priori knowledge using both a shape-based vector and a dynamic text-based vector discovering significant words throughout the process.

In section II, we introduce the A2ING and some related systems. Hence, we describe our system enhancement in section III and comment our results in section IV. Perspectives are given as conclusion in section V.

II. Related works

Incremental classification is not a new method in machine learning. Basically, incremental classification learns along the process to cover the samples representation space according to given descriptors. Its benefits are plural. As it learns along the process, it does not need all necessary classes at the beginning and thus, can discover new data. When a class has enough elements to classify documents, it stops asking the class label to the user [9]. So, we can reduce the number of sample for each class.

Polikar et al.[14] gives an overview of several algorithms for incremental classification. Some of them are an evolution of classical algorithms (incremental SVM [12], Incremental K-means [15]). Other approaches are based on Incremental Growing Neural Gas (IGNG) and variations like the one proposed by Hamza & al.[6] for clustering (unsupervised classification). Among IGNG family, Bouguelia & al.[9][10] have proposed an incremental semi-supervised classifier (A2ING). This system is introduced below (Figure 1).

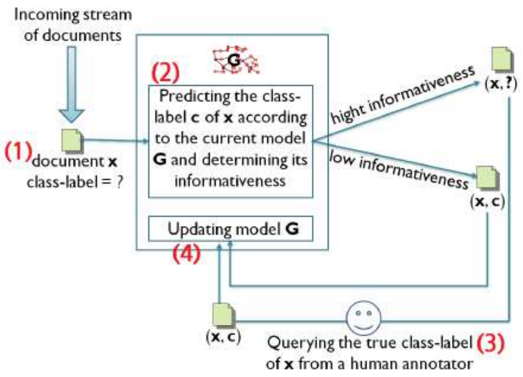
Incoming stream
of documents
(1)
document x
class-label = ?
(2)
G
Predicting the class-label c of x according to the current model G and determining its informativeness
hight informativeness
(x, ?)
low informativeness
(x, c)
Updating model G
(4)
(x, c)
Querying the true class-label
(3)
of x from a human annotator

Figure 1. Rafik Bouguelia's A2ING general scheme[10].

(1) Each object X is represented as a vector. When an unknown object X is classified (2), according to an informativeness criteria, the object is rejected (3) for user annotation or predicted to a class C. During the annotation (3) the user can give an existing class reinforcing (4) the system or create dynamically a new class to expend the system scope (4).

Theoretically, the system is never ending. It adapts itself due to the ability of introducing a new class on real-time.

Drawbacks, all these different algorithms rely on a set of fixed descriptors to feature all classes, usually, a vector X representing the object to classify. The performance comes from both the classifier and the feature vector to figure out discriminations. In all the quoted references, different feature vectors are used showing the flexibility of the classifier to various problems. Bouguelia [9] experimented normalized snippet pixels to feature characters, temporal and spatial features for on-line characters, bag-of-words for textual documents, etc. In any case, the vector is defined a priori, fixed size and hence, it closes obviously the ability to feature a new problem. For instance, when using a bag-of-words which does not include English words, you cannot classify a document written in English.

Literature offers many descriptor proposals oriented for the classification purpose: for detecting human being in a scene, we can use gradients, colours information [11]; for handwriting/machine print classification, we can measure linearity and profile regularity on pseudo words [16]; for logo classification or document classification, we can use a Blurred Shape Model [3] or pixel density quantification on patches [4]; for structure document classification, layout feature or structural features can be used [2]; LLAH approaches for retrieving textual documents [7]; and weighted bag-of-words for text classification [5]. Bag-of-words are usually large vectors of words where each word defines a dimension.

To summarize all these experiences, graphical or pixel based descriptors are fine for structured information (forms, logo, tables, etc.) but less efficient in case of variable document structures (news, emails, etc.). Text-based descriptors are more robust to the document structure variation, allowing natural language text classification. These descriptors are words or group of words supposed to be discriminant for the classification. To select this lexicon, a previous statistical analysis of the domain is required. Basically, Term Frequency (TF) and Inverse Document Frequency (IDF) methods [5] can be applied.

According to us, the definition of the feature vector for the incremental classification of textual images is a bottleneck. Moreover, we cannot plan neither the domain nor the structure of any new image class appearing along our process.

Hamza and al.[6] and Bouguelia and al.[10] successfully experimented textual features (bag-of-words) for textual documents incremental classification. However, they never discussed the lexicon selection because this topic was not in the center of their works. They suppose a preprocessing stage to define the vector before to start the incremental learning. Here, a prior TF analysis on text samples was performed to select words for the vector. This preprocessed word learning stage is contradictory with our incremental classification objectives. An additional difficulty is that the feature vector proposed for the A2ING shall be fixed-size. This is due to the used vector distance function (cosine or euclidian). Notice the similar issue for many standard classifiers (SVM, Perceptron…). But obviously, the bag-of-words dimension depends on the variability and complexity of the corpus vocabulary. It differs from a domain to another. [5] demonstrates on different corpus (Popol and Reuters) the impact of the vector dimension on classification results. A generic model cannot be a fixed-size vector.

Deep Learning approaches [17][18] offer today a strategy to avoid the explicit feature selection. For instance, quoted reference authors apply deep learning for text understanding from character-level inputs. They use temporal convolutional networks to let the system discovers relevant features. Even if this approach is interesting, it is not yet compatible for incremental learning based on few samples discovered along the process.

As described above, image analysis approaches provide generic structural and holistic descriptors but they are inefficient and difficult to tune [4] for poorly structured documents. Unfortunately, they often appear in our image workflow (receipts, invoices, bank notice, payslips, etc.)

To cope with this challenge, we design an innovative and real full incremental system for textual document images.

III. PROPOSAL

We propose to combine a standard shape-based descriptor and an original adaptive and generic textual descriptor with the A2ING.

A. *Shape-based classification*

Many shape-based descriptors are compliant to document incremental classifiers because they can be fully computed during the process, they are fixed-size, and they are independent of the document semantic or the document content (genericity). For all these reason, we propose to use a shape-based vector. One of them is the Blurred Shape Model (BSM)[3]. The BSM splits the picture in 8x8 squares and each square is calculating from a blur pixel representation. We have chosen this method due the simplicity and the demonstrated efficiency on various document image classification problems: structured document classification, and logo classification [8].

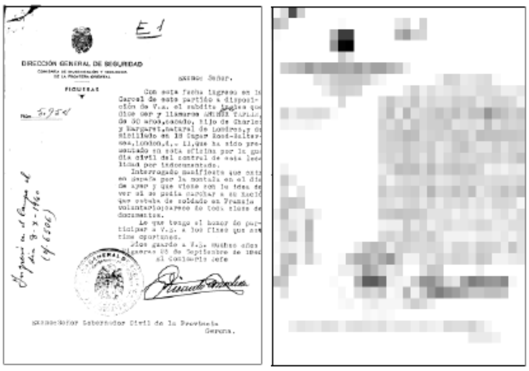

Figure 2. Example of blurred image by the BSM

But, this method is not efficient on semi-structured textual document because in this case the information is more carried by the text and words than the structure itself.

B. Text-based classification

Hence, a text-based vector for textual classification is an alternative method when the classification depends on accurate semantic textual information. Secondly, we can have documents (text in a natural scene) without recurrent structural information.

In the initial A2ING, the text-based vector is pre-computed. Our proposal is to enhance the system by dynamically discovering and expending the dimension of the text-based vector along the system life.

At the beginning, the feature vector X of the A2ING is empty (no dimension). Then, the vector X_i with i dimension is enlarged by n dimensions ($X_{i+n} = X_i \cup w_n$) when a set w_n of n “relevant” words are discovered to model a new class. This definition is compatible with an Euclidian distance because w_n new features are valued to 0 for any existing classes (already modelled by the X_i vector). At this stage, any new classes will be modelled by the X_{i+n} vector.

The issue is both to discover the w_n features and to decide when the text-based vector dimension shall be enlarged.

The discovery of relevant words for text classification is not a new topic. There are several statistical metrics in the state-of-the-art to figure out the relevance of words in a corpus. The most well-known method [5] is based on the Term Frequency / Inverse Document Frequency (TF/IDF). Basically, selected words are those frequent for one class and not frequent in all the others. It means selected words are discriminative. A more semantic method named Latent Semantic Analysis is used to make correlations between words in a document. It produces topic features instead of word features. Finally, these methods calculate a weight of a word or a concept to rank them. The selection is given by the top weights according to the ranking.

These statistic approaches need a representative training set to model several variables: language, domain, classes, etc. In our case, all these issues cannot be pre-defined because we do not know anything about captured images.

Bouillot [5] demonstrates that there is no best solution for all the cases but the only metric which is both dependent of an image class and independent of others is the Term Frequency. Far to be the best metric, TF is interesting because it can be computed incrementally each times a new sample appears without constraint from other classes and future unknown classes.

In this study, we propose the Term Frequency as a first approach for the w_n evaluation.

Most m frequent terms $W_m(k)$ for a class k give some key recurrent features for an image sample of the class. Statistically, we except $W_m(k)$ to be included (at least partially) within the terms of the image related to k. Let suppose we have j dimensions for the feature vector X_j, the vector T_j is the terms of an image limited to the identified X_j features and $X_j(k)$ is the A2ING instanced model vector for a class k:

the distance $D(k)=T_{j-} X_j(k)$ is minimized for the class k when $W_m(k)$ is included in X_j because $W_m(k)$ should be included in T_j as explained above. Then, the A2ING can predict this class.

If $W_m(k)$ is not included in $X_{j,}$ the distance D(k) is maximized and the informativeness criteria of the A2ING will reject the prediction. In this case $X_{j+m} = X_j \cup W_m(k)$ will minimize D for the class k and in the same time may maximize D for any other class. The prediction is enhanced. The feature vector is dynamically enlarged.

The issue is to select the n terms w_n among $W_m(k)$ to add to the vector X. In this first study we propose to limit w_n to the n words with the best TF value for the class k and which are not yet in X when an image prediction is rejected. Actually, we set n to 1 to get the most frequent term not yet in X representing k. But it may happen that few terms occur always together due to an equal TF. With this strategy, we introduce a minimum number of “best” terms. If added terms are sufficient and discriminative to predict the class, then the incremental classification is optimal, otherwise the system will wait for further samples. The system will manage itself up to a sufficient number of terms to predict a class.

What happens if the system can never learn an image class and the X vector increase as infinite? This could be dramatic, moreover if image class samples occur frequently. To be honest, we have not yet deal with this question which is a perspective. For the moment we threshold the system to a maximum number of M considering that if the system cannot learn an image class with M Terms means the class is unpredictable.

Another difficult question is to decide when $W_m(k)$ is relevant. If only few images were captured for a class k, W(k) is not representative. Waiting for more samples to take a decision will delay the incremental learning, by keeping X out of W(k) inputs. This question is still to be explored. We have not yet found out a solution and we work around with a parameter giving a minimum number of samples to threshold the TF. This parameter can be set by experiments.

C. *Multi-classifier-based classification*

We can describe our system by the figure 3 below.

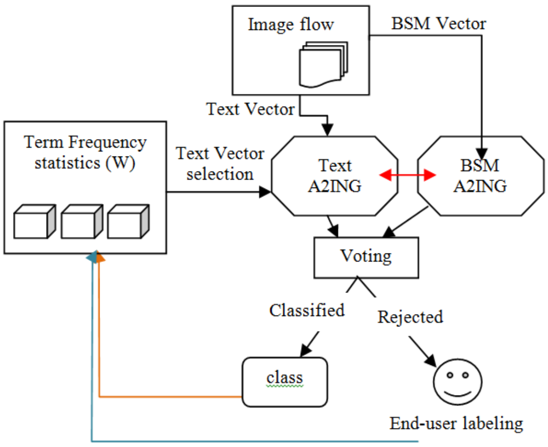

Figure 3. Schema of classification process

The BSM vector is computed directly on the image. The text classifier needs an OCR processing to compute the Text Vector X. The combination of both is a cooperative-concurrent classifiers.

1) *Concurrent classifiers*

Each classifier is an A2ING based classifier allowing parallel incremental learning. Each one has its own feature vectors: BSM Vector and Text Vector.

2) *Cooperative classifiers*

Both A2ING deliver a class prediction. Each answer (red link in figure 3) can be used as a feedback for the other to learn without waiting for the end-user feedback. End-user feedback or classification success enables the TF computation for the Text Vector selection (if needed).

IV. Result Comparison

We have experimented our proposal on the ITESOFT corpus that Bouguelia used for its measures. We describe this corpus below. Reusing this corpus, we can compare our proposal to the “a priori” defined textual vector as in [9][10].

A. *Dataset*

The ITESOFT corpus is available on demand according a NDA. Images are machine-printed document like invoices, mails, forms, etc. We have both TIF images and OCR readings. Thus, we can easily use shape and text vectors.

The corpus includes two datasets so-called MMA and LIRMM. MMA contains 2591 images divided on 25 classes. LIRMM contains 1951 images and 24 classes. To compare with Bouguelia, the dataset is splitted in two parts:

- Learning phase: to initialize a classifier and do not measure from scratch because the lack of samples biaised the measure. We take 2/3 of the dataset.

- Test phase: measure on remaining documents (1/3 of the corpus) while the classifier continues to learn.

B. *Results on different approaches*

We evaluate each single A2ING and the combination (table 1). We compare to initial Bouguelia works (table 2). For the quantification of the classification performance, we used the standard Recall and Precision measures like in [9].

TABLE I. Classification results with different approaches

	LIRMM dataset		**MMA dataset**	
	Recall (%)	***Precision (%)***	***Recall (%)***	***Precision (%)***
BSM	57	98	31	96
Text	78	96	62	96
Multi	82	98	66	93

BSM has the worst performance because the corpus contains mainly semi-structured documents. This vector is efficient only on very structured images. Hence the performance of the Multi-classifier really comes from the Text-classifier. However, BSM is useful for a part of the dataset: few classes with structured images and very few instances of image. In this case, the TF does not reach our threshold to be learnt by the Text-classifier.
Result difference between LIRMM and MMA is explained by different corpus complexity (more variability [9]).

C. *Comparison with previous work.*

Results “[9]” for the comparison between the non-generic vector A2ING and our proposal come from Bouguelia report [9].

TABLE II. Results with our proposal and Bouguelia approach

	LIRMM dataset		**MMA dataset**	
	Recall (%)	***Precision (%)***	***Recall (%)***	***Precision (%)***
Proposal	82	98	66	93
[9]	95,2	95,6	75	78,4

Unquestionably, recall is much better with Bouguelia approach while the precision is better with our system. However, performances are quite acceptable when you consider that the system started from scratch. It demonstrates both that descriptors can be learnt incrementally and that an A2ING can cope with a growing feature vector.

Our analysis shows two important reasons of the reduced performance:

- First comes for the selection of the vector X. Only based on the TF metric, iteratively updated by image along the processing of each dataset, we cannot converge to the same dictionary than a pre-computed TF/IDF. The table III demonstrates the difference. In our proposal we retrieve more than 91% of the TF/IDF dictionary from [9]. This is very good but we introduce a lot of unexpected additional words. They are for instance named entities (first name, last name, city names, etc.) or unrelevant words (natural langage syntactic operator like "like" "you", "of", "the"…) occurring in many images. For MMA the text variability is larger hence many best frequent words are not so frequents even if they are still the best frequent. The impact is to increase the D distance and in consequence the rejection. Positively it impacts the precision but it is a side effect.

TABLE III. FEATURE VECTOR SIZE COMPARISON

	LIRMM dataset	MMA dataset
Features vector in [9]	277	292
Our features vector	341	1700
Rate of features in [9] included in our	91%	93%

- A second reason is the learning delay of our approach. First learnt classes are re-learnt during the process because the first learnt classes are only based on a small X vector. The increasing of selected features maximizes the distance with previous learnt classes when a learnt class shares some new introduced descriptors. Fortunately, the system manages itself the relearning but introduces a delay in the network convergence and of course, more feedback from the user. Notice that 10% recall difference is 10% more rejection and hence, 10% more user feedbacks. In perspective, we plan to evaluate larger corpus to analyse this issue.

V. CONCLUSION AND FUTURE WORK

All these observations show the importance of the feature selection criteria. TF seems an interesting proposal because independent of other classes but not yet sufficient to filter unexpected terms. For instance generic terms shared by different classes are filtered by the IDF. The exploration of the criteria enhancement is a major perspective, like simulating the IDF or exploring the TF standard deviation.

The system was set-up for text vectors. However, our statistic approaches to discover a feature and embed it into A2ING is generic for any kind to feature which we can be observed within images. Our principle of a full incremental system for image classification could help computer vision and robotics to adapt to different progressive environments.

In conclusion we demonstrate both that we can have a full incremental efficient system, starting from scratch with really no prior knowledge and that an A2IGN can cope with a dynamic incremental feature vector. This system gives acceptable performance and several perspectives exist.

REFERENCES

[1] A. Joseph, "Automatic Detection of Fixed Expressions" PhD report, Université Paris 13, 2013

[2] A. Antonacopoulos, C. Clausner, C. Papadopoulos, and S. Pletschacher, "Icdar 2013 competition on historical newspaper layout analysis" IEEE International Conference on Document Analysis and Recognition, pp. 1454-1458, 2013

[3] A. Fornés, S. Escalera, J. Lladós, G. Sánchez, and J. Más, "Hand Drawn Symbol Recognition by Blurred Shape Model Descriptor and Multiclass Classifier" in "Graphics Recognition. Recent Advances and New Opportunities" Lecture Notes in Computer Science, vol.5046, pp. 30-40, Springer-Berlag, Berlin. 2008

[4] F. Alaei, N. Girard, S. Barrat, and JY. Ramel, "A New One-class Classification Method Based on Symbolic Representation: Application to Document Classification" 11th IAPR International Workshop on Document Analysis Systems, Tours, France, pp. 00-00, 2014

[5] F. Bouillot, "Text Classification: new weights adapted for small samples" PhD report, university of Montpellier, 2015

[6] H. Hamza, Y. Belaïd, A. Belaïd, and B. Baran Chaudhuri, "Incremental classification of invoice documents" 19th IEEE International Conference on Pattern Recognition (ICPR), Dec 2008, Tampa, United States, 2008

[7] K. Takeda, K. Kise, and M. Iwamura, "Real-Time Document Image Retrieval for a 10 Million Pages Database with a Memory Efficient andStability Improved LLAH" IEEE International Conference on Document Analysis and Recognition (ICDAR), Beijing, pp. 1054 - 1058, 2011

[8] M. Rusiñol, V. Poulain d'Andecy, D. Karatzas, and J. Llados, "Classification of Administrative Document Images by Logo Identification" GREC 2011, Seoul, Korea, Volume 7423 of the series Lecture Notes in Computer Science, pp. 49-58, 2011

[9] M-R. Bouguelia, "Classification and Active Learning from dynamic dataflows with incertain labels" PhD report, Université de Lorraine, 2015

[10] M-R. Bouguelia, Y. Belaïd, and A. Belaïd, "A stream-based semi-supervised active learning approach for document classification" IEEE International Conference on Document Analysis and Recognition (ICDAR), Washington DC (USA), pp. 611-615, August 2013

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection" '05 Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pp. 886-893, 2005

[12] P. Laskov, C. Gehl, K. Stefan, and K-R. Müller, 2006. "Incremental Support Vector Learning: Analysis, Implementation and Applications" J. Mach. Learn. Res. 7 (December 2006), pp. 1909-1936, 2006

[13] P. Sidiropoulos; S. Vrochidis; and I. Kompatsiaris. "Adaptive hierarchical density histogram for complex binary image retrieval" International workshop on Content-based Multimedia Indexing (CBMI), 2010

[14] R. Polikar, L. Upda, S. S. Upda, and V. Honavar. 2001. "Learn++: an incremental learning algorithm for supervised neural networks" Trans. Sys. Man Cyber Part C 31, 4, pp. 497-508, november 2001

[15] S. Chakraborty, N.K. Nagwani, and L. Dey. "Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms" International Journal of Computer Applications 27, pp. 14-18, August 2011

[16] S. Hamrouni, F. Cloppet, and N. Vincent, "Handwritten and printed text separation: linearity and regularity assessment" International Conference Image Analysis and Recognition, ICIAR14, Vilamoura, Portugal, pp. 387-394, 2014

[17] X. Zhang and Y. LeCun, "Text Understanding from Scratch" Technical Reports eprint arXiv : 1502.01710, february 2015

[18] X. Zhang, J. Zhao, and Y. LeCun, "Character-level Convolutional Networks for Text Classification" Advances in Neural Information Processing Systems 28 (NIPS), arXiv:1509.01626, december 2015