# Classification active de flux de documents avec identification des nouvelles classes

Mohamed-Rafik Bouguelia, Yolande Belaïd, Abdel Belaïd

# Classification active de flux de documents avec identification des nouvelles classes

**Mohamed-Rafik Bouguelia — Yolande Belaïd — Abdel Belaïd**

***Université de Lorraine*** *- LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-**54506**, France*
*Email: {mohamed.bouguelia, yolande.belaid, abdel.belaid}@loria.fr*

RÉSUMÉ. *Dans cet article, on propose un algorithme semi-supervisé actif pour la classification de flux continu de documents. Cet algorithme, basé sur une méthode adaptative d'apprentissage non supervisé, permet de repérer les documents les plus informatifs à l'aide d'une mesure d'incertitude pour demander leur étiquette à un opérateur. Il construit et maintient un modèle sous forme d'un graphe à topologie dynamique dont les nœuds sont des représentants de documents étiquetés, formant ce qu'on appelle l'"espace couvert par les classes connues". Il permet de détecter automatiquement les nouvelles classes apparaissant dans le flux. Un document est identifié comme membre d'une nouvelle classe ou d'une classe connue, selon qu'il se trouve à l'extérieur ou à l'intérieur de l'espace couvert par les classes connues. Les expérimentations effectuées sur des ensembles de documents réels montrent que la méthode nécessite peu de documents à étiqueter et qu'elle atteint des performances comparables aux méthodes supervisées qui sont entraînées sur des ensembles de documents présents en mémoire et entièrement étiquetés.*

ABSTRACT. *In this paper, we propose a stream-based semi-supervised active learning method for document classification, which is able to query (from an operator) the class labels of documents that are informative, according to an uncertainty measure. The method maintains a dynamically evolving graph topology of labelled document-representatives, which constitutes a covered feature space. The method is able to automatically discover the emergence of novel classes in the stream. An incoming document is identified as a member of a novel class or an existing class, depending on whether it is outside or inside the area covered by the known classes. Experiments on different real datasets show that the proposed method requires a small amount of the incoming documents to be labelled, in order to learn a model which achieves better or equal accuracy than to the usual supervised methods with fully labelled training documents.*

MOTS-CLÉS : *Flux de documents, Classification, Identification de nouvelles classes*

KEYWORDS: *Document stream, Classification, Novel class identification*

CIFED 2014, pp. 75–89, Nancy, 18-21 mars 2014

## 1. Introduction

Les administrations traitent des milliers de documents administratifs hétérogènes qui sont quotidiennement numérisés. Ces documents numérisés doivent être traités rapidement et efficacement afin de permettre un gain de productivité pour les administrations et un gain de réactivité et de temps de réponse pour les utilisateurs. Ils doivent être classés dans différentes catégories telles que des chèques bancaires, des reçus de frais médicaux, des factures, des ordonnances, etc. L'objectif est de les diriger automatiquement vers des dispositifs de traitements spécifiques (selon leur catégorie) ou vers des services spécialisés dans leur gestion. La classification de ces documents par thème permet également de donner du contexte afin de lever des ambiguïtés concernant la reconnaissance de certains mots clés pour l'analyse du document. Pour certaines applications réelles, les méthodes de classification de documents usuelles comme celles décrites dans (N. Chen, 2007) ne sont pas adaptées, pour différentes raisons détaillées ci-après.

Afin d'obtenir un bon taux de reconnaissance, les méthodes usuelles doivent être entraînées en utilisant de nombreux documents étiquetés. Beaucoup d'entre elles sont donc entièrement supervisées et coûteuses étant donné qu'elles nécessitent l'étiquetage manuel d'un grand nombre de documents. Les techniques d'apprentissage semi-supervisées (Zhu, 2008 ; M. Zaki, 2008) peuvent apprendre en utilisant des données étiquetées et non étiquetées, et peuvent donc être utilisées pour alléger le coût d'étiquetage. Les documents à étiqueter peuvent être sélectionnés au hasard ou choisis par l'algorithme de classification. Il s'agit alors d'"apprentissage actif" (S. Ertekin, 2007 ; Y. Fu, 2012 ; A. Bordes, 2005). C'est l'algorithme qui décide de demander les étiquettes de classes de certains documents "importants" à un opérateur.

Les méthodes existantes de classification de documents ont besoin de connaître à l'avance l'ensemble des documents utilisés pour l'apprentissage. Ceci n'est pas possible dans le cas d'un flux où les documents arrivent en continu et deviennent disponibles au fil du temps (flux infini). Dans ce cas, l'apprentissage se fait au fur et à mesure que les documents arrivent : chaque nouveau document est utilisé immédiatement et une seule fois, pour mettre à jour le modèle de façon incrémentale.

La plupart des méthodes de classification incrémentales pour les flux de données (MM. Gaber, 2007 ; J. Kolter, 2005) supposent généralement que le nombre de classes de documents dans le flux est connu. Cependant, dans les applications réelles, cette supposition est souvent incorrecte. En effet, il est difficile d'obtenir des échantillons de documents étiquetés à partir de toutes les classes de documents possibles qui apparaitront dans le futur. De plus, la nature évolutive du flux fait que de nouvelles classes de documents peuvent apparaître à tout moment. Si une nouvelle classe de document n'est pas automatiquement détectée, toutes les instances de documents de cette classe seront inévitablement classées par erreur dans des classes existantes (connues). L'identification de nouvelles classes est donc importante pour éviter de telles erreurs de classification. Les méthodes d'apprentissage actif à partir de flux de documents (Y. Fu, 2012 ; A.B. Goldberg, 2011) supposent implicitement que les documents uti-

lisés pour leur initialisation, couvrent toutes les classes possibles, et elles demandent à un opérateur les étiquettes des documents dont l'incertitude est déterminée selon ces classes connues. Par conséquent, elles ne parviennent pas à détecter les nouvelles classes.

Nous avons proposé dans (M.R. Bouguelia, 2013) une méthode d'apprentissage non supervisée incrémentale qui peut apprendre continuellement à partir d'un flux continu de documents. Cette méthode est peu sensible aux paramètres d'initialisation. Dans cet article, nous présentons une extension semi-supervisée active de cette méthode. Cette extension permet également de détecter automatiquement les nouvelles classes de documents qui peuvent apparaître dans le flux. Elle peut apprendre de façon incrémentale à partir d'un flux continu de documents, et maintient ainsi une topologie dynamique (graphe) de nœuds qui constituent des représentants de documents. Les nœuds sont les centres d'hypersphères qui couvrent des régions de l'espace où des documents de classes connues ont été observés. Un document qui se trouve en dehors de la zone couverte est considéré comme appartenant à une classe nouvelle. La méthode proposée permet de demander seulement les étiquettes des documents les plus informatifs : ceux qui sont incertains selon une mesure décrite dans la section 4.1, et ceux qui sont considérés comme appartenant à des classes nouvelles.

Le reste de l'article est organisé comme suit. Dans la Section 2, nous évoquons l'état de l'art et nous situons notre méthode par rapport aux méthodes existantes. Dans la Section 3, nous décrivons la méthode non supervisée (AING) qui est à l'origine de la méthode proposée. Dans la Section 4, nous étendons cette méthode à un apprentissage semi-supervisé actif pour la classification des documents avec détection de nouvelles classes. Une évaluation expérimentale sur des documents réels est détaillée dans la Section 5. Dans la Section 6, nous concluons et présentons quelques perspectives de ce travail.

## 2. Travaux connexes

Quelques méthodes incrémentales et semi-supervisées pour la classification de documents textes sont présentées dans (C.C. Aggarwal, 2012), mais ces méthodes ne sont pas actives et ne permettent donc pas de sélectionner les documents informatifs à étiqueter manuellement. Des méthodes actives et semi-supervisées pour la classification des flux de données, sont disponibles dans la littérature (Y. Fu, 2012 ; A.B. Goldberg, 2011) mais elles ne permettent pas de prendre en compte de nouvelles classes.

Les méthodes traditionnelles de détection de nouveautés telles que les techniques de détection de données aberrantes et les SVM à une classe (B. Scholkopf, 2001 ; Agarwal, 2005 ; S. Subramaniam, 2006) supposent qu'il n y a qu'une seule classe de données "normales" et que toute donnée qui dévie de cette classe est une donnée "anormale" ou "nouvelle". De plus, ces méthodes ne sont pas capables d'apprendre à partir d'un flux continu. Quelques méthodes comme (E.J. Spinosa, 2008) sont basées sur la classification non supervisée pour la détection de nou-

velles classes dans un flux de données, mais elles supposent également qu'il n y a qu'une seule classe de données connue et que toutes les autres classes sont nouvelles. Quelques méthodes actives comme celles présentées dans (J. He, 2007 ; Escudeiro, 2012) offrent des stratégies pour découvrir des classes inconnues, mais elles ont besoin que l'ensemble des données soit disponible à l'avance, ce qui n'est pas concevable pour un *flux* de documents.

Les techniques de détection et suivi de topiques, dites TDT (Topic Detection and Tracking), permettent principalement de détecter et de suivre de nouveaux événements dans un flux de données (e.g. la détection d'informations liées à partir de différentes sources de presse). Les événements détectés ne constituent pas forcément de nouvelles classes de données. Le terme "topique" employé par ces méthodes, désigne un événement unique qui se produit à un moment donné dans le temps (Allan, 2002). C'est cet événement ainsi que sa réapparition et son évolution qui sont détectés par ces méthodes.

La méthode que nous proposons est semi-supervisée active et destinée à la classification des flux de données infini. Elle est différente des méthodes existantes car elle est capable d'identifier l'apparition de nouvelles classes dans le flux, contrairement aux méthodes traditionnelles de détection de nouveautés.

## 3. Apprentissage non supervisé avec AING

Nous avons proposé une première méthode appelée AING : "Adaptive Incremental Neural Gas". L'algorithme de cette méthode construit et maintient un modèle sous forme d'un graphe $G$ (Figure 1) dont les nœuds sont des représentants de documents. Chaque nœud $y \in G$ est un vecteur de caractéristiques qui est mis à jour en permanence par l'algorithme.

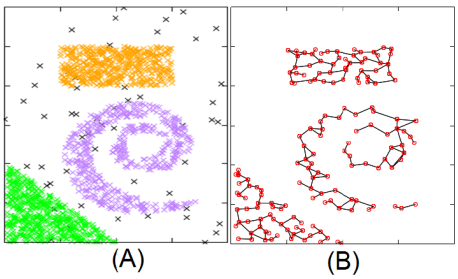


**Figure 1.** *Apprentissage de la topologie des données. (A) : données synthétiques en 2 dimensions. (B) : graphe G de représentants.*

Soit $x \in \mathbb{R}^p$ le vecteur de caractéristiques représentant un nouveau document [1]. On considère que $x$ est *assez loin* (respectivement *assez proche*) d'un nœud $y$ si la distance entre $x$ et $y$ est supérieure (respectivement inférieure) à un seuil $s$. Le principe général de l'algorithme peut être exprimé selon 3 cas.

Soit $y_1$ et $y_2$ respectivement les $1^{er}$ et $2^{eme}$ nœuds les plus proches de $x$, telle que $\text{dist}(x, y_1) < \text{dist}(x, y_2)$. L'algorithme met à jour le modèle $G$ à chaque nouvelle arrivée d'un document, comme suit :

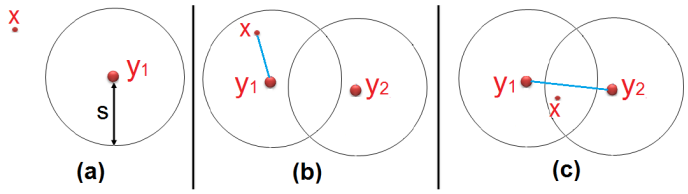


**Figure 2.** *Les 3 cas de l'algorithme pour la génération et l'adaptation des nœuds du graphe $G$*

1) **si** $\text{dist}(x, y_1) > s$ **alors** // (Figure 2(a))

- $G \leftarrow G \cup \{y_{new} | y_{new} = x\}$ // nouveau nœud $y_{new}$ en se basant sur $x$.

2) **si** $\text{dist}(x, y_1) < s$ et $\text{dist}(x, y_2) > s$ **alors** // (Figure 2(b))

- $G \leftarrow G \cup \{y_{new} | y_{new} = x\}$
- Relier $y_{new}$ à $y_1$ par une nouvelle arête.

3) **si** $\text{dist}(x, y_1) < s$ et $\text{dist}(x, y_2) < s$ **alors** // (Figure 2(c))

- $x$ est attribué à $y_1$
- Relier $y_1$ à $y_2$ par une nouvelle arête (si $y_1$ et $y_2$ ne sont pas déjà reliés)
- $y_1 \leftarrow y_1 + \epsilon_1 \times (x - y_1)$ // modifier $y_1$ pour être moins distant de $x$.
- $\forall y_i \in N_{y_1} : y_i \leftarrow y_i + \epsilon_2 \times (x - y_i)$ // modifier le vecteur de chaque nœud $y_i$ voisin de $y_1$ ($N_{y_1}$ est l'ensemble de tous les nœuds reliés à $y_1$ par une arrête) pour être moins distant de $x$.

Notez que la distance que nous utilisons dans cet article est la distance Euclidienne. Cependant, toute autre mesure peut être utilisée, e.g., la mesure cosinus qui peut parfois être plus adaptée lorsqu'il s'agit de sac-de-mots de très grande taille.

Quand un document $x$ est assez proche de ses deux plus proches nœuds $y_1$ et $y_2$, il est attribué à $y_1$ ($3^{eme}$ cas). Ce dernier ainsi que ses nœuds voisins sont mis à jour (ils se rapprochent de $x$) par un taux d'apprentissage : $\epsilon_1$ pour $y_1$ et $\epsilon_2$ pour ses nœuds voisins. Comme discuté dans (M.R. Bouguelia, 2013), un taux d'apprentissage trop grande implique l'instabilité des nœuds, tandis qu'un taux d'apprentissage trop petit

1. Les caractéristiques dépendent de l'application. Nous utilisons dans les expérimentations une représentation du document sous forme de sac-de-mots textuels, afin de les classer par thèmes.

implique que les nœuds n'apprennent pas assez à partir des documents qui leur sont attribués. Les valeurs caractéristiques sont $0 < \epsilon_1 \ll 1$ et $0 < \epsilon_2 \ll \epsilon_1$. Soit $n_{y_1}$ le nombre de documents attribués à $y_1$. Dans AING, $\epsilon_1 = \frac{1}{n_{y_1}}$, il diminue lentement, proportionnellement au nombre de documents attribués à $y_1$ (plus $y_1$ apprend, plus il devient stable), et $\epsilon_2$ est tout simplement défini comme $\epsilon_2 = \frac{1}{n_{y_1} \times 100}$, c'est-à-dire 100 fois plus petite que la valeur de $\epsilon_1$ ($\epsilon_2 \ll \epsilon_1$).

Notez que la méthode est incrémentale et n'as pas besoin de sauvegarder les documents déjà vus. La graphe maintenu ($G$) évolue dynamiquement en fonction des nouveaux documents.

## 4. Extension en apprentissage semi-supervisé actif

L'algorithme précédant est non supervisé et n'est pas directement applicable à la tâche de classification souhaitée. Nous étendons cette méthode pour apprendre à partir de documents étiquetés et non étiquetés. Au lieu de choisir aléatoirement les documents à étiqueter manuellement à partir du flux, nous laissons l'algorithme décider à chaque nouvelle arrivée d'un document, si oui ou non l'étiquette de sa classe doit être demandée à un opérateur.

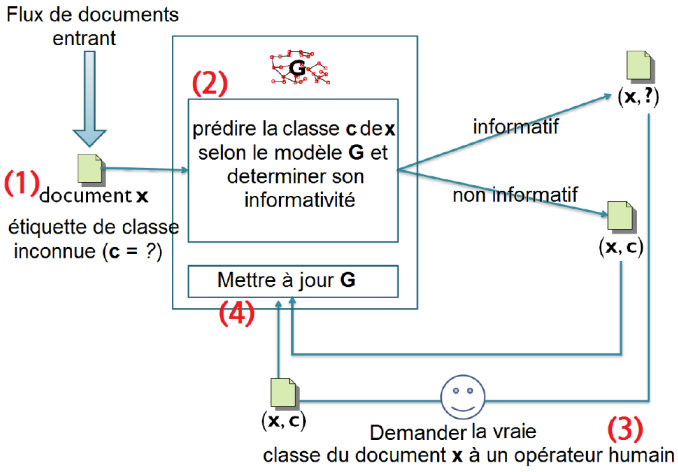


**Figure 3.** *Le schéma général de la méthode*

Nous utilisons les premiers documents entrants du flux pour initialiser le modèle $G$ avec quelques nœuds[2]. Ensuite, chaque nouveau document représenté par un vecteur de caractéristiques $x$ (Figure 3 (1)) est classé dans une classe $c$ selon le modèle actuel

2. Dans nos expérimentations, nous avons initialisé le modèle avec seulement 1% des documents, choisis aléatoirement à partir de l'ensemble de documents.

$G$ (Figure 3 (2)). Un un indice d'incertitude est calculé pour déterminer si le document $x$ est informatif. La méthode utilisée est décrite dans les sections 4.1 et 4.2. Si le document est considéré comme informatif, sa vraie classe (étiquette) $c$ est demandée à un opérateur (Figure 3 (3)). Le document classé $(x, c)$ est alors appris (Figure 3 (4)) pour mettre à jour et améliorer le modèle $G$ tel que décrit dans la section 4.3.

### 4.1. *Classification et informativité d'un document*

Pour chaque nouveau document $x$, nous utilisons la méthode KNN (L. Jiang, 2007) pour dériver une probabilité d'appartenance à ses deux classes les plus probables.

Soit $\mathrm{KNN}(x) = \{(y_1, c_{y_1}), ..., (y_K, c_{y_K})\}$ les $K$ nœuds les plus proches de $x$ sélectionnés à partir de $G$, triés dans l'ordre croissant en fonction de leur distance à $x$. $P(c|x)$, la probabilité que le document $x$ appartient à la classe $c$, est déterminée comme suit :

$$P(c|x) = \frac{\sum_{(y_i, c_{y_i}) \in \mathrm{KNN}(x)} f(y_i, c_{y_i})}{K} \quad [1]$$

où $f(y_i, c_{y_i}) = \begin{cases} 1 & \text{si } c_{y_i} = c \\ 0 & \text{sinon} \end{cases}$

Pour un document $x$ donné, notons $c_1 = \underset{c}{\operatorname{argmax}}\, P(c|x)$ et $c_2 = \underset{c \neq c_1}{\operatorname{argmax}}\, P(c|x)$ respectivement les première et seconde classes les plus probables, telle que $P(c_1|x) \geq P(c_2|x)$.

Soit $\Delta_{(c_1, c_2|x)} = P(c_1|x) - P(c_2|x)$. Un document avec une petite valeur $\Delta$ est plus incertain, parce que la probabilité d'appartenir à sa classe la plus probable $c_1$ est proche de la probabilité d'appartenir à sa deuxième classe la plus probable $c_2$. Autrement dit, plus $\Delta_{(c_1, c_2|x)}$ est proche de 0 plus la classe du document $x$ est incertaine et plus le document $x$ est informatif, car connaître la vraie classe d'un tel document serait utile au modèle $G$ afin de mieux discriminer entre ces classes (Voir Section 4.3).

Pour décider si la classe d'un nouveau document $x$ doit être demandée à un opérateur (c'est-à-dire si $x$ est informatif), nous définissons une valeur de confiance $\delta$. Si $\Delta_{(c_1, c_2|x)} < \delta$ alors la vraie classe de $x$ est demandée à un opérateur. Sinon, le document $x$ est classé comme $c_1$ (sa classe prédite la plus probable).

Pour illustrer intuitivement quels sont les documents incertains (avec une faible valeur $\Delta$) qui sont informatifs pour notre modèle, la Figure 4 (a) montre trois classes de données synthétiques qui se chevauchent, et la Figure 4 (b) montre les données pour lesquelles on demande la classe, c'est-à-dire celles qui sont situées dans une région d'incertitude et donc considérées comme informatives. Connaître leur véritable étiquette de classe permettra de mieux séparer les classes qui se chevauchent.

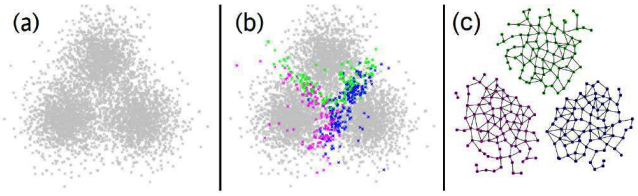


**Figure 4.** *(a) 3 classes de données synthétiques (suivant une distribution gaussienne) qui se chevauchent. (b) Les données de classes incertaines. (c) La topologie des nœuds obtenue après avoir supprimé les arêtes inter-classes.*

Notez que nous pouvons rendre la valeur de confiance $\delta$ adaptative. Supposons que $\Delta_{(c_1,c_2|x)} < \delta$. Dans ce cas, la vraie classe de $x$ est demandée à l'opérateur. Notons cette véritable classe par $c_x^*$. Si la classe la plus probable $c_1$ a été correctement prédite (c'est à dire $c_1 = c_x^*$), alors on peut être plus confiant et donc légèrement diminuer la valeur de $\delta$. Sinon, si la classe la plus probable $c_1$ n'était pas la vraie classe (c'est à dire $c_1 \neq c_x^*$), alors on peut être moins confiant et augmenter donc légèrement la valeur de $\delta$.

### 4.2. *Espace couvert et détection de nouvelles classes*

La stratégie précédente se focalise sur la demande des étiquettes des documents incertains dont l'incertitude est déterminée en fonction des classes connues ($\Delta_x$ mesure l'ambiguïté entre les deux classes existantes les plus probables). Par conséquent, comme les méthodes actives usuelles, cette stratégie n'explore pas les régions de l'espace où de nouvelles classes peuvent apparaître et échoue donc à les détecter, sauf si par hasard, des documents de nouvelles classes se trouvent dans la région d'incertitude.

**Définition** : *Soit un seuil de distance noté $s$. Chaque nœud $n_i \in G$ constitue le centre d'une hypersphère définie par le rayon $s$. L'union de tous les hypersphères est appelée la "zone couverte de l'espace des caractéristiques". Un document $x$ est à l'extérieur (resp. intérieur) de la zone couverte si la distance à son plus proche nœud $y_1 \in G$ est supérieure (resp. inférieure) à $s$.*

Un document se trouvant à l'extérieur de la zone couverte, est un membre d'une classe nouvelle, contrairement à un document qui se trouve à l'intérieur de l'espace couvert. Si $s$ est trop petit, de nombreux documents seront considérés à tort comme des instances de nouvelles classes. De même, lorsque $s$ est trop grand, de nombreux documents seront considérés à tort comme des instances de classes existantes. En réalité, il est très difficile d'initialiser manuellement $s$ car sa valeur dépend fortement des données (l'ensemble des documents). Par conséquent, étant donné que notre algorithme est semi-supervisé et actif, nous utilisons l'étiquette associée à chaque document pour ajuster automatiquement la valeur de $s$.

Soit $x$ un document dont la classe (étiquette) a été prédite ou demandée à l'opérateur (i.e. si $x$ est incertain ou en dehors de la zone couverte). Soit $y_1$ le nœud le plus proche de $x$ et $\epsilon$ une constante de petite valeur tel que $0 < \epsilon \ll 1$. Le seuil (i.e. rayon) $s$ est adapté de façon incrémentale comme suit :

– Si $x$ est en dehors de la zone couverte : si étiquette$(x)$ = étiquette$(y_1)$ alors le fait de déclarer $x$ comme membre d'un classe nouvelle, est considéré comme faux. Dans ce cas, on augmente la valeur de $s$ comme suit :

$$s := s + \epsilon \times |\operatorname{dist}(x, y_1) - s|$$

– Si $x$ est à l'intérieur de la zone couverte : si étiquette$(x) \neq$ étiquette$(y_1)$ alors le fait de déclarer $x$ comme membre d'une classe existante, est considéré comme faux. Dans ce cas, on diminue la valeur de $s$ comme suit :

$$s := s - \epsilon \times |\operatorname{dist}(x, y_1) - s|$$

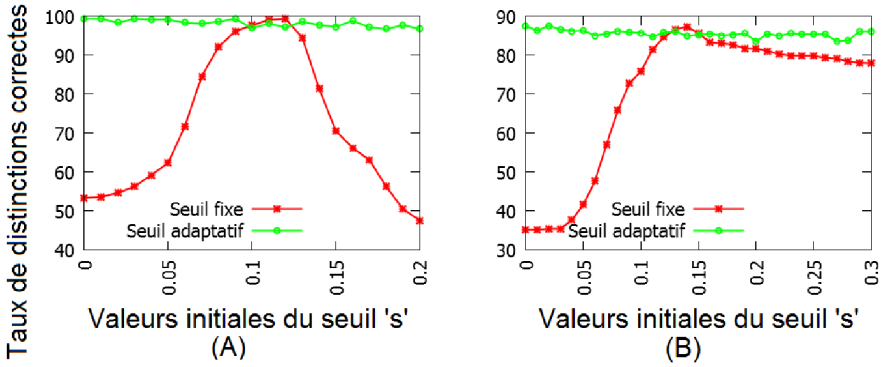


**Figure 5.** *La capacité à distinguer correctement les documents appartenant à des classes existantes de ceux qui constituent de nouvelles classes, selon différentes valeurs du seuil s, sur deux ensembles de documents différents*

Afin de montrer la capacité de la méthode proposée à détecter correctement les documents appartenant à de nouvelles classes grâce à la stratégie adaptative décrite dans cette section, la Figure 5 montre les taux de distinction entre les classes nouvelles et existantes selon différentes valeurs de $s$, sur deux ensembles de documents : "docSet" et "LIRMM" (voir la section expérimentations). Lorsqu'une valeur fixe de $s$ est utilisée (c'est-à-dire, seuil fixé manuellement et non adaptatif), nous observons une valeur optimale autour de $r = 0.12$ pour l'ensemble de documents "docSet" (voir Figure 5.A) et $r = 0.14$ pour l'ensemble de documents "LIRMM" (voir Figure 5.B). Tandis que lorsque la valeur de $s$ est adaptative, le résultat obtenu est toujours proche de l'optimal quel que soit la valeur initiale de $s$. Ceci prouve que la méthode est insensible aux valeurs initiales du seuil $s$, qui peut donc toujours être initialisé à $s = 0$ (i.e. la plus petite valeur initiale possible).

### 4.3. *Mise à jour du modèle* $G$

Soit $(x, c_x)$ une nouvelle donnée où $x$ est le vecteur de caractéristiques du document et $c_x$ sa classe prédite ou demandée (comme décrit dans la Section 4.1 et 4.2). La mise à jour du modèle $G$ telle qu'elle est présentée dans AING est modifiée uniquement pour le $3^{ème}$ cas de l'algorithme décrit dans la Section 3. L'adaptation proposée a pour but de maximiser la séparation entre les différentes classes qui se chevauchent.

Soit $y_1$ et $y_2$ respectivement les $1^{er}$ et $2^{ème}$ nœuds les plus proches de $x$, telle que $\text{dist}(x, y_1) < \text{dist}(x, y_2)$ :

**Si** $\text{dist}(x, y_1) < s$ et $\text{dist}(x, y_2) < s$ **alors**

– Relier $y_1$ à $y_2$ par une arête (s'ils ne sont pas déjà reliés)

– **Si** $c_x = c_{y_1}$ :

- $y_1 \leftarrow y_1 + \epsilon_1 \times (x - y_1)$
- $\forall y_i \in N_{y_1}$ et $c_x \neq c_{y_i} : y_i \leftarrow y_i - \epsilon_2 \times (x - y_i)$

– **Si** $c_x \neq c_{y_1}$ :

- $y_1 \leftarrow y_1 - \epsilon_1 \times (x - y_1)$
- $\forall y_i \in N_{y_1}$ et $c_x = c_{y_i} : y_i \leftarrow y_i + \epsilon_2 \times (x - y_i)$

Lorsque $x$ est assez proche de $y_1$ et de $y_2$, le nœud $y_2$ devient un voisin de $y_1$ (i.e. $y_1$ et $y_2$ sont liés par une arête). Si ces deux nœuds sont labellisés différemment, l'arête les reliant sera une arête inter-classes. L'algorithme peut déterminer une meilleure séparation entre les classes, en éloignant deux nœuds voisins étiquetés différemment. En effet, si $y_1$ a la même étiquette que $x$ (i.e. $c_x = c_{y_1}$), alors $y_1$ est mis à jour pour être moins éloigné de $x$, et les nœuds voisins de $y_1$ qui sont étiquetés différemment de $x$, sont mis à jour pour être plus éloignés de $x$. D'autre part, si $y_1$ est étiqueté différemment de $x$ (i.e. $c_x \neq c_{y_1}$), alors $y_1$ est mis à jour pour être plus éloigné de $x$, et les nœuds voisins de $y_1$ qui sont étiquetés comme $x$ sont mis à jour pour être moins éloignés de $x$. La Figure 4 (c) montre la topologie obtenue après avoir enlevé les arrêtes inter-classes, pour les 3 classes de la Figure 4 (a).

## 5. Évaluation expérimentale

Nous considérons dans notre évaluation expérimentale différents ensembles de documents administratifs réels fournis par des clients de la société ITESOFT[3]. Chaque document est d'abord traité par un OCR et représenté par un sac-de-mots textuels, un vecteur qui contient les nombres d'occurrences des mots dans le document. Les vecteurs de chaque ensemble sont représentés dans un espace à $p$ dimensions, où $p$ est la

3. http ://www.itesoft.com

taille du vocabulaire [4]. Les ensembles de documents testés contiennent des nombres différents de classes (de 13 à 141 classes) :

– **DocSet** : 779 documents, $p = 413$, 13 classes.

– **CAF** : 1158 documents, $p = 271$, 141 classes.

– **LIRMM** : 1951 documents, $p = 277$, 24 classes.

– **MMA** : 2591 documents, $p = 292$, 25 classes.

– **APP** : 19742 documents, $p = 328$, 139 classes.

Nous testons, en plus de la méthode proposée, une variante incrémentale active de SVM ("LASVM") (A. Bordes, 2005). Les données les plus proches de la limite de décision de SVM sont celles qui sont les plus susceptibles d'être étiquetées manuellement. Nous utilisons aussi pour effectuer des comparaisons un certain nombre de classifieurs classiques (KNN (L. Jiang, 2007), LogitBoost (J. Friedman, 1998), NaiveBayes (D. Lowd, 2005) et RandomForest (Breiman, 2001)). Pour rappel, la méthode que nous proposons offre deux qualités supplémentaires comparée aux classifieurs considérés (sauf LASVM) : (i) seulement quelques documents considérés informatifs sont manuellement labellisés durant le processus de classification, (ii) la classification et l'apprentissage sont faits en parallèle en traitant les documents un par un à la volée (à la différence des méthodes classiques où chaque document peut être revisité de nombreuses fois au cours de l'apprentissage). Les mesures d'évaluation considérées sont : le taux de documents qui ont été manuellement étiquetés, le taux d'erreurs, ainsi que la précision et le rappel moyens.

Pour une classe $c$ donnée, soit $vp_c$ le nombre de documents correctement classés dans la classe $c$ (vrais positifs), $fp_c$ le nombre de documents incorrectement classés dans la classe $c$ (faux positifs), et $fn_c$ le nombre de documents qui ont pour vrai classe $c$, mais n'ont pas été classés dans la classe $c$ (faux négatifs). La précision de la classe $c$ est $precision_c = \frac{vp_c}{vp_c+fp_c}$, elle exprime le ratio de documents correctement classés dans la classe $c$ par rapport au nombre total de documents classés dans la classe $c$. La rappel de la classe $c$ est $rappel_c = \frac{vp_c}{vp_c+fn_c}$, il exprime le ratio de documents correctement classés dans la classe $c$ par rapport au nombre total de documents de la classe $c$. Nous utilisons le rappel et précision moyens par rapport à toutes les classes. Le taux d'erreurs est exprimé comme le nombre total de documents incorrectement classés sur le nombre de documents.

Les résultats obtenus sont présentés dans le Tableau 1. Tout d'abord, pour notre méthode, le nombre de documents qui ont été étiquetés manuellement (en interrogeant un opérateur) représentent en moyenne 36,3 % du nombre total de documents utilisés, ce qui est meilleur que le taux obtenu par LASVM. Les autres méthodes, sont entièrement supervisées et ont donc besoin que l'ensemble des documents utilisés pour l'apprentissage soit étiqueté. Le Tableau 1 montre que pour l'ensemble de documents *DocSet* et *CAF*, notre méthode réalise les meilleures performances avec LogitBoost pour *DocSet* et NaiveBayes pour *CAF*. En ce qui concerne l'ensemble

4. Ceci est le nombre de mots significatifs ou fréquents pour chaque ensemble de documents

| **Méthode** | **Étiquettes %** | **Erreur %** | **Précision %** | **Rappel %** |
|---|---|---|---|---|
| Ensemble DocSet | | | | |
| **Méthode proposée** | **39.1%** | **19.2** | 80.9 | 80.7 |
| KNN | tous | 25.7 | 76.0 | 74.2 |
| **LogitBoost** | tous | **19.2** | 80.9 | **80.8** |
| NaiveBayes | tous | 25.3 | 76.6 | 74.6 |
| RandomForest | all | 21.1 | 77.9 | 78.8 |
| LASVM | 51.05% | 20.7 | **81.3** | 79.2 |
| Ensemble CAF | | | | |
| **Méthode proposée** | **53.8%** | **28.7** | 75.7 | **71.2** |
| KNN | tous | 31.6 | 74.7 | 68.4 |
| LogitBoost | tous | 38.0 | 69.9 | 61.9 |
| **NaiveBayes** | tous | **28.7** | **75.8** | **71.2** |
| RandomForest | tous | 32.6 | 72.5 | 67.4 |
| LASVM | 66.9% | 29.2 | 73.9 | 70.7 |
| Ensemble LIRMM | | | | |
| Méthode proposée | **22.5%** | 3.8 | 96.1 | 96.1 |
| KNN | tous | 5.6 | 94.6 | 94.3 |
| LogitBoost | tous | 4.4 | 95.4 | 95.5 |
| NaiveBayes | tous | 4.4 | 96.1 | 95.5 |
| **RandomForest** | tous | **3.07** | **96.9** | **96.8** |
| LASVM | 42.2% | 3.53 | 96.2 | 96.4 |
| Ensemble MMA | | | | |
| Méthode proposée | **39.7%** | 23.8 | 79.0 | 76.1 |
| KNN | tous | 27.2 | 76 | 72.8 |
| LogitBoost | tous | 27.4 | 73.1 | 72.5 |
| NaiveBayes | tous | 24.4 | 76.5 | 75.6 |
| RandomForest | tous | 25.7 | 76 | 74.3 |
| **LASVM** | 43.5% | **22.5** | **79.1** | **77.4** |
| Ensemble APP | | | | |
| **Méthode proposée** | **26.4%** | **14.6** | 85.3 | **85.3** |
| KNN | all | 15.8 | 84.4 | 84.2 |
| LogitBoost | tous | 18.9 | 81.6 | 81.0 |
| NaiveBayes | tous | 22.8 | 81.1 | 77.1 |
| RandomForest | tous | 16.0 | 84.3 | 83.9 |
| LASVM | 30.2% | 15.22 | **85.7** | 84.8 |
| Résultats moyens sur tous les ensembles de documents | | | | |
| **Méthode proposée** | **36.3%** | **18.02** | **83.4** | **81.88** |
| KNN | tous | 22.26 | 81.14 | 78.78 |
| LogitBoost | tous | 21.58 | 80.18 | 78.34 |
| NaiveBayes | tous | 21.12 | 81.22 | 78.8 |
| RandomForest | tous | 19.69 | 81.25 | 80.24 |
| LASVM | 46.77% | 18.23 | 83.24 | 81.7 |

**Tableau 1.** *Résultats expérimentaux sur différents ensembles de documents*

*LIRMM*, bien que la méthode proposée exige que seulement $22,5\%$ des documents soient étiquetés, elle réalise une meilleure performance que KNN, LogitBoost et NaiveBayes. Cependant, RandomForest et LASVM réalisent une meilleure performance que notre méthode. Pour l'ensemble *MMA*, LASVM est légèrement supérieur à notre méthode, tandis que pour l'ensemble *APP*, notre méthode réalise les meilleures performances en termes de taux d'erreurs et de rappel. Enfin, les résultats moyens sur tous les ensembles de documents sont indiquées en bas du Tableau 1. Nous pouvons constater que la méthode proposée atteint, en moyenne, les meilleures performances en ce qui concerne le taux d'erreur, la précision et le rappel, tout en utilisant moins de documents étiquetés que les autre méthodes.

La Figure 6 (côté gauche) indique, pour chaque ensemble de documents, la précision obtenue par rapport à l'effort humain qui est exprimé en termes de nombre de documents qui sont étiquetés manuellement. Figure 6 (à droite) montre le nombre de documents étiquetés par rapport au nombre de documents parcourus dans le flux. La méthode proposée est comparée à LASVM. Le cas "passif" sur la Figure 6 représente le cas où chaque document du flux est manuellement étiqueté et utilisé pour l'apprentissage même s'il n'est pas informatif. Pour tous les ensembles de documents, nous pouvons voir que la méthode proposée atteint un taux de reconnaissance meilleure ou égale à LASVM en demandant uniquement les étiquettes des documents les plus informatifs, et atteint toujours un meilleur résultat que le mode passif qui perd du temps à demander les étiquettes de classes de tous les documents, qu'ils soient informatifs ou non.

## 6. Conclusion et travaux futurs

En se basant sur AING, nous avons présenté dans ce papier une méthode efficace pour la classification de flux de documents. La méthode proposée est adaptée à un environnement industriel où les documents deviennent disponibles progressivement au fil du temps, alors que leurs étiquettes ne sont pas disponibles. En effet, l'algorithme : (i) peut apprendre en ligne à partir d'un flux de documents arrivant en continu, (ii) ne demande que les étiquettes des documents les plus informatifs, et économise ainsi le temps et l'effort qu'un humain doit fournir lors de l'étiquetage, et (iii) est capable de détecter les documents de nouvelles classes qui peuvent apparaître dans le flux. La méthode ne nécessite pas que le nombre de classes ou de nœuds soit connus. Outre le fait que la méthode proposée est incrémentale et qu'elle réduit considérablement le nombre de documents à étiqueter, les résultats expérimentaux sur des ensembles de documents réels montrent son efficacité par rapport à des classifieurs entièrement supervisés et actifs.

Dans le futur, nous envisageons d'améliorer cette méthode en introduisant une notion de "coût d'étiquetage" des documents. En fonction du type et de la qualité des documents, on peut prendre en compte un coût d'étiquetage variable selon les documents. Ce coût peut être, par exemple, fonction de la durée moyenne nécessaire pour étiqueter chaque type de document.

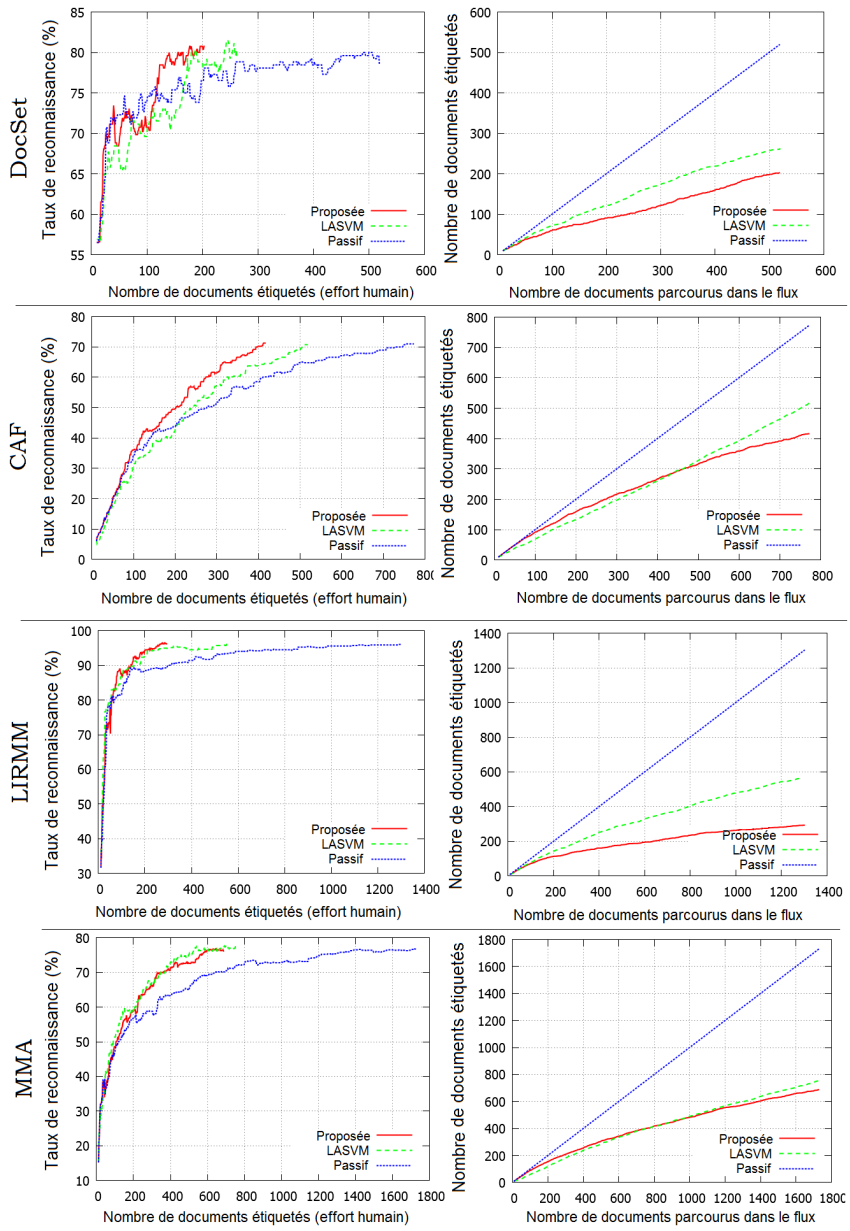


**Figure 6.** *Gauche : taux de reconnaissance selon le nombre de documents étiquetés. Droite : Nombre de documents étiquetés selon le nombre total de documents reçus à partir du flux*

## 7. Bibliographie

A. Bordes S. Ertekin J. W. L. B., « Fast kernel classifiers with online and active learning », *JMLR,* p. 1579-1619, 2005.

A.B. Goldberg e. a., « OASIS : Online Active Semi-Supervised Learning », *AAAI,* p. 1-6, 2011.

Agarwal D., « An Empirical Bayes Approach to Detect Anomalies in Dynamic Multidimensional Arrays », *ICDM,* p. 1-8, 2005.

Allan J., « Introduction to topic detection and tracking », *Topic detection and tracking. Springer US,* p. 1-16, 2002.

B. Scholkopf J. Platt J. S.-T. A. J. S. R. C. W., « Estimating the support of a high-dimensional distribution », *Neural Computation,* p. 1443-1471, 2001.

Breiman L., « Random Forests », *Machine Learning,* p. 5-32, 2001.

C.C. Aggarwal C. Z., « A survey of text clustering algorithms », *Mining Text Data. Springer US,* p. 77-128, 2012.

D. Lowd P. D., « Naive Bayes models for probability estimation », *ICML,* p. 529-536, 2005.

E.J. Spinosa A.P. de Leon F. d. C. J. G., « Cluster-Based Novel Concept Detection in Data Streams Applied to Intrusion Detection in Computer Networks », *VLDB,* p. 976-980, 2008.

Escudeiro N. F. J. A. M., « D-Confidence : an active learning strategy to reduce label disclosure complexity in the presence of imbalanced class distributions », *JBCS,* p. 311-330, 2012.

J. Friedman T. Hastie R. T., « Additive Logistic Regression : a Statistical View of Boosting », *Annals of Statistics,* p. 337-407, 1998.

J. He J., « Nearest-neighbor-based active learning for rare category detection », *NIPS,* p. 633-640, 2007.

J. Kolter M. M., « Using Additive Expert Ensembles to Cope with Concept Drift », *ICML,* p. 449-456, 2005.

L. Jiang Z. Cai D. W. S. J., « Survey of improving k-nearest-neighbor for classification », *FSKD,* p. 679-683, 2007.

M. Zaki H. Y., « Semi-supervised Growing Neural Gas for Face Recognition », *JMMA,* p. 425-435, 2008.

MM. Gaber A. Zaslavsky S. K., « A survey of classification methods in data streams », *In Data Streams - Springer US,* p. 39-59, 2007.

M.R. Bouguelia Y. B. A. B., « An adaptive incremental clustering method based on the growing neural gas algorithm », *ICPRAM,* p. 1-8, 2013.

N. Chen D. B., « A survey of document image classification : problem statement, classifier architecture and performance evaluation », *IJDAR,* p. 1-16, 2007.

S. Ertekin J. Huang L. B. C. G., « Learning on the Border : Active Learning in Imbalanced Data Classification », *CIKM,* p. 127-136, 2007.

S. Subramaniam T. Palpanas D. P. V. K. D. G., « Online Outlier Detection in Sensor Data Using Non-Parametric Models », *VLDB,* p. 187-198, 2006.

Y. Fu X. Zhu B. L., « A survey on instance selection for active learning », *KAIS,* p. 1-35, 2012.

Zhu X., « Semi-supervised learning literature survey », *Computer Sciences Technical Report 1530, University of Wisconsin-Madison,* , 2008.