# Exploration des mémoires à court et long terme pour la classification multi-labels en flux

Xihui Wang[*,**], Pascale Kuntz[**], Frank Meyer[*]

[*]Orange Labs, 2 Avenue Pierre Marzin, 22300 Lannion
prenom.nom@orange.com,
[**]Laboratoire des Sciences du Numérique de Nantes - Site Polytech - 44300 Nantes
prenom.nom@univ-nantes.fr

**Résumé.** La classification multi-labels, dans laquelle un texte, une image ou une cyber-attaque, par exemple, peuvent être associés à plusieurs labels simultanément devient de plus en plus nécessaire dans les applications récentes. Lorsque les besoins de réactivité sont eux-mêmes cruciaux, la classification de flux de données devient un enjeu important. Nous proposons dans cet article un nouvel algorithme, Online Memory k-means (OMk), pour traiter la problématique de la classification en flux. OMk est un modèle de type k-plus-proches-voisins qui utilise deux types de mémoire, l'une court-terme basée sur une fenêtre glissante FIFO, et l'autre long-terme, basée sur un échantillonnage en réservoir. Ces deux mémoires permettent de gérer les flux de données avec des dérives de concepts et de pouvoir résister au phénomène d'oubli catastrophique. En utilisant ces structures de données simples avec des tailles de mémoire relativement limitées et en considérant l'information portée par les corrélations entre labels, notre algorithme est compétitif avec les algorithmes actuels de l'état de l'art, EaHTps et MLSAMPKNN, à la fois en qualité de prédiction et en temps de réponse. La faible complexité d'OMk et les performances obtenues nous permettent d'envisager son extension à la classification multi-labels extrême de données en flux, qui est un problème encore peu exploré.

## 1 Introduction

Le paradigme de la classification multi-labels est une extension du problème mono-label à plusieurs labels qui sont extraits d'un ensemble prédéfini de possibilités : il s'agit d'associer un objet décrit par un vecteur de variables à un sous-ensemble restreint de concepts d'intérêt, appelés "labels". Pour illustration, en annotation de textes, on peut qualifier un texte à la fois de "drôle", "en français" et "consacré au cinéma". La classification multi-labels a connu un fort essor cette dernière décennie stimulée par de nombreuses applications : par exemple, en biologie pour la classification des fonctions des gènes et des protéines (Clare et King, 2001), en science de l'information pour la catégorisation de textes (Kong et Philip, 2011a) ou en multimédia pour la catégorisation et l'annotation d'images, de vidéos et de musiques (Madjarov et al., 2012). Depuis les travaux pionniers de Boutell et al. (2004), Tsoumakas et al. (2009)