

Stratégies de classification non supervisée basées sur fenêtres superposées : application aux données d'usage du Web

Alzenny Da Silva, Yves Lechevallier

Projet AxIS, INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay cedex – France
{Alzenny.Da_Silva, Yves.Lechevallier}@inria.fr
<http://www-rocq.inria.fr/axis>

Résumé. Un problème majeur se pose dans le domaine des flux de données : la distribution sous-jacente des données peut changer sur le temps. Dans cet article, nous proposons trois stratégies de classification non supervisée basée sur des fenêtres superposées. Notre objectif est de pouvoir repérer ces changements dans le temps. Notre approche est appliquée sur un benchmark de données réelles et les conclusions obtenues sont basées sur deux indices de comparaison de partitions.

1 Introduction

Dans cet article, nous proposons trois stratégies de classification non supervisée appliquées sur fenêtres superposées. Notre objectif est de pouvoir repérer les changements de la distribution sous-jacente d'un flux de données sur le temps. Notre approche consiste donc à fixer *a priori* la taille de la fenêtre et appliquer un algorithme de classification non supervisée sur les données contenues à l'intérieur de la fenêtre. Nous définissons deux types de partitionnement de données sur les fenêtres : partitionnement par nombre d'effectifs (fenêtre logique) et partitionnement par intervalle de temps (fenêtre de temps).

L'idée principale est de faire glisser la fenêtre sur le temps de telle façon que des nouvelles données soient rajoutées dans la fenêtre et par conséquent, les données les plus anciennes en soient éliminées. L'action de glissement de la fenêtre sur les données est fait de telle manière à ce qu'il y ait toujours une zone de chevauchement entre les deux ensembles de données contenues dans la fenêtre avant et après son glissement. Chaque fois qu'une nouvelle fenêtre est définie, l'algorithme de classification non supervisée est appliqué sur les données contenues dans la fenêtre, ce qui définit une partition et un ensemble de prototypes. La détection des possibles changements est faite par la comparaison de deux partitions obtenues sur le même ensemble d'individus. Dans ce contexte, nous proposons trois types de comparaisons de partitions : comparaison sur les données de l'intersection, comparaison sur les données de l'union et comparaison sur la totalité des données.