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

Alzennyr Da Silva, Yves Lechevallier

Projet AxIS, INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 Le Chesnay cedex – France
{Alzennyr.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 donnés 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 conséquence, 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 supervisé 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.

2 Algorithme et critères d'évaluation

Pour la classification de données dans notre approche, nous avons utilisé l'algorithme K-means (MacQueen, 1967) et la distance Euclidienne pour le calcul de dissimilarité entre deux individus. Comme étude de cas, nous utilisons un jeu de données d'usage du Web issu d'une entreprise polonaise et diffusé dans le cadre du challenge ECML/PKDD 2007 [1]. Pour analyser les résultats, nous utilisons deux critères : la F-mesure (van Rijsbergen, 1979) et l'indice de Rand corrigé (Hubert et Arabie, 1985).

3 Conclusion

Comme résultat de toutes expérimentations, nous avons des valeurs très élevées (compris entre 0.8 et 1) pour les deux critères d'évaluation, ce qui nous montre un jeu de données assez stable et sans changements remarquables. Cela peut être dû à la courte période de temps disponible pour l'analyse : pas plus de 22 jours, les premiers de l'année. En conclusion, il est évident la nécessité d'établissement d'un compromis entre la taille de la fenêtre et le temps de réponse désiré, mais aussi en considérant la périodicité des changements.

Comme prolongement de ces expérimentations, nous citons l'application de cette approche sur d'autres jeux de données - aussi bien réelles qu'artificielles - et aussi l'exécution d'autres simulations afin d'analyser l'influence des différentes valeurs des paramètres d'entrée (tels que le pourcentage de chevauchement et le nombre de clusters). De plus, nous envisageons la mise en place de dispositifs permettant la découverte automatique de la périodicité présente dans le flux de données.

Références

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*, 193–218.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *5th Berkley Symposium on Mathematics and Probability*, Volume 1, pp. 281–297.

van Rijsbergen, C. J. (1979). *Information Retrieval* (second ed.). London : Butterworths.

Summary

A major difficulty is present in the data stream domain: the underlying data distribution may change over time. In this article, we propose three strategies of unsupervised classification based on overlapping windows. Our aim is to detect these changes over time. Our approach is applied on a benchmark of real data. The conclusions obtained are based on statistical analysis.

[1] Plus de détails à propos de ce challenge sur : `http://challenge.ecmlpkdd2007.org/challenge/`