Présentation des travaux de thèse

Antoine GHORRA

Département informatique et automatique- Institut Mines Telecom

24/05/2017

Plan de la présentation

- Introduction
- Etat de l'art
  - Méthodologies de Clustering
  - Méthodologies de classification
- Présentation du projet d'article de review de l'état de l'art.
- Etudes et résultats sur DenStream
- Présentation de l'algorithme modifié
- Perspectives et travaux futurs.

Introduction

Suite aux travaux effectués au sein de l'équipe de recherche du département DIA viens le sujet de la thèse en question concernant le développement de méthodologies pour le classification/clustering des données de manière incrémentale et en ligne.

- Pour effectuer une étude incrémentale et en ligne deux manières se présentent

Classification	Clustering
Apprentissage supervisé	Apprentissage non-supervisé
Notions de classes	Notions de Clusters

Table 1: Quelques différences entre Classification et Clustering

Etat de l'art

Article Review Etat de l'art

Après étude de l'état de l'art concernant les différents aspects de classification et/ou clustering qui peuvent être utilisés, on a songé à l'écriture d'un article de review de l'état de l'art qui sera présenté dans les deux semaines qui arrivent.

Base de données

La base de données utilisée dans les articles traitant les différentes méthodologies de clustering est KDD CUP 99'.

Etat de l'art

Facteur	DenStream	Clustream	D-Stream	Grid-based DBSCAN
Flot continu	Oui	Oui	Oui	Oui
Nombre de Clusters connu	Non	Oui		
Cluster de tailles aléatoire	Oui	Non	Oui	Non
Modèle à deux phases		Oui		Oui
Traitement de données bruitées	Oui	Non	Oui	Oui

Table 2: Etude comparative des différentes méthodes de clustering en ligne incrémentale existante dans l'état de l'art

DenStream

DenStream est une méthodologie permettant de faire la détection des clusters dans un flot de données continue.
DenStream est basé sur les concepts suivants:

- La notion des micro-clusters pour prendre en compte les clusters de taille arbitraire.
- Une stratégie d'élimination des données inutiles.
- La présentation de ressources mémoire pour les outliers afin de bien gérer la mémoire.
- Qualité de clustering élevée.

DenStream

Plusieurs paramètres sont pris en considération lors du calcul mathématique pour DenStream:

Paramètre	Fonction
λ	Facteur de dégradation temporelle
μ	Seuil du nombre des éléments d'un micro-cluster
β	Facteur de pondération
ϵ	Rayon des micro-clusters

Table 3: Les différets paramètres de l'algorithme DenStream et leur fonctionnemment

Dans l'article initial présentant DenStream les vleurs des paramètres en question sont: $\lambda = 0.25$, $\beta = 0.2$ $\mu = 0.25$ $\epsilon = 16$

Expérimentations

Theorem

Les paramètres choisis ne représentent pas les valeurs optimales pour obtenir le résultat le plus proche du nombre de classes dans le benchmark.
le nombre de classes dans le benchmark est de 23.

Proof.

Mettre le tableau ici pour montrer les combinaisons avec les valeurs les plus proches du nombre de classes en question et avec les valeurs de la redondance de meme et faire une étude comparative avec les paramètres présentés dans l'article meme. □

Modifications proposées

Definition

L'importance des données est un facteur majeur décidant l'élimination des clusters au sein de l'algorithme.

Theorem

Comsidérer la densité des données dans un micro-cluster comme facteur essentiel de dégradation des données.

- Récupérer tous les micro-clusters.
- Calculer les coordonnées du centre du cluster.
- Calculer le nombre des élements dans le micro-cluster
- Récupérer tous les élements des micro-clusters
- Calculer la distance du centre par rapport a chaque point
- Considerer la plus grande distance comme étant rayon de la sphere.
- Calculer le volume de la sphère puis diviser ce volume par le nombre des élements pour obtenir la densité.

Soit $X_1, X_2, \ldots, X_n$ est l'ensemble des élements du micro-cluster X_c le centre du micro-cluster. d La valeur de la distance entre le centre et les différents élements du micro-cluster calculée de la manière suivante:
$d = \sqrt{(x_c - x_i)^2 + (y_c - y_i)^2}$
N étant le nombre des éléments d'un micro cluster $i \in [1, N]$
$r = max(d)$ correspond au rayon de la sphère comme distance maximale
Le volume de l sphère se calcul par $V = \frac{4 * \pi * r^3}{3}$
Finalement la densité sera calculée par $D = \frac{V}{N}$
Deux sorties possibles des modifications de l'algorithme:

1. Remplacement du facteur temporel par la densité.
2. Faire une combinaison entre les deux facteurs.