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

Expérimentations

DenStream est une méthodologie permettant de faire la détection des clusters dans un flot de données continue.

Readable Mathematics

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n} \sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.