Etude de la scabilité des méthodes d'optimisation d'architecture des réseaux de neurones. Application à l'estimation des redshifts photométriques

Nina Pelagie BEKONO

LIMOS, Université Blaise Pascal

09 Avril 2015

Plan

sommaire

Galaxie

Définition

Assemblage d'étoiles, de gaz, de poussières et de matière noire, contenant parfois un trou noir supermassif en son centre. Exemple: La voie lactée.

Différents Types: elliptique, Spirale, Irrégulière.

Figure : Capture de galaxies

Redshift Photométrique

Definition

Phénomène astronomique de décalage vers les grandes longueurs d'ondes *(rouge)* des raies spectrales et de l'ensemble du spectre d'une galaxie.

- **Cause:** la dilatation de l'espace provoquée par l'expansion de l'univers.

Redshift Photométrique

Definition

Phénomène astronomique de décalage vers les grandes longueurs d'ondes *(rouge)* des raies spectrales et de l'ensemble du spectre d'une galaxie.

- **Cause:** la dilatation de l'espace provoquée par l'expansion de l'univers.

Interêt

Evaluer la distance d'une galaxie par rapport à un observateur

- Redshift (z) => Vitesse d'éloignement ($v = cz$).
- $V = H_0 D$ (Relation d'Hubble).

Methodes Spectroscopiques

1. Observer les galaxies avec un élément disperseur pour séparer les différentes longueurs d'onde de la lumière.
2. Identifier des raies d'émission et d'absorption.
3. Déduire le décalage spectral.

Caractéristiques

- Efficaces.
- Précises.
- Coûteuses en temps.

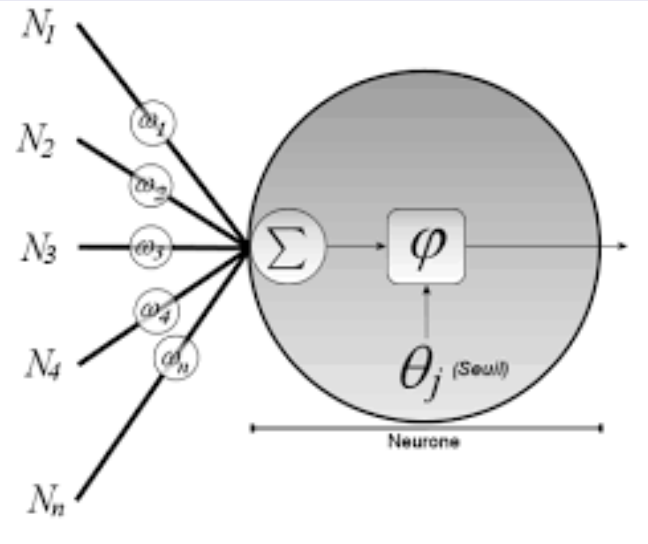
Methodes Photométriques

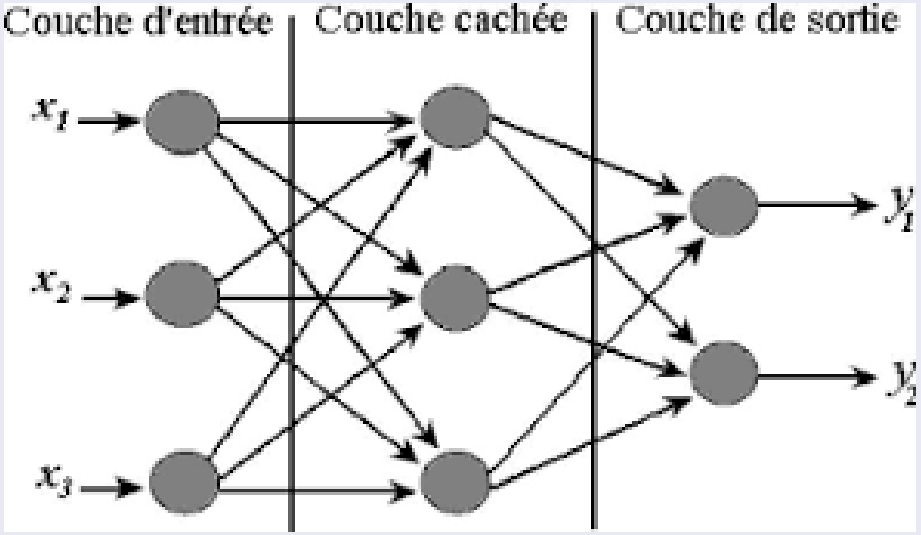
Utilisation des filtres relativement grossiers.

1. Template Fitting ou ajustement des SED aux modèles.
2. Methodes empiriques: Utiliser le Redshift pour calibrer un algorithme d'apprentissage.
   Plusieurs techniques:
   - Les arbre de decision.
   - Support Vector Machines.
   - Forêts aléatoires.
   - Réseaux de neurones.

Définition

Les reseaux de neurones artificiels





Neurone formel:

- Combinaison linéaire des entrées: $v = \sum_{i=1}^{n} w_i * x_i$ avec w_i les poids, x_i les entrées et n le nombre d'entrées.
- Sortie $y = f(v)$ avec f la fonction d'activation.

Problème

RNA: Approximateurs universels.

Trouver l'architecture adaptée

1. quel est le nombre de couches du reseau?
2. quel est le nombre de neurones par couche?
3. comment les connecter?
4. comment initialiser des poids de connexions?

Géneralement le choix est fait par l'utilisateur.

Problème

RNA: Approximateurs universels.

Trouver l'architecture adaptée

1. quel est le nombre de couches du reseau?
2. quel est le nombre de neurones par couche?
3. comment les connecter?
4. comment initialiser des poids de connexions?

Géneralement le choix est fait par l'utilisateur.
Deux approches d'optimisation dans la littérature:

- soit de manière dynamique: *MTLING, MTower...*
- soit de manière statique: *treillis de Galois,...*

Estimation de Redshifts photométriques => Problème de regression. Des travaux existants: *Collister et Lahav 2004, M.Brescia et al. 2012,F.Feroz et al. 2014 ...* mais pas d'optimisation d'architecture de réseaux de neuronnes.

Estimation de Redshifts photométriques => Problème de regression. Des travaux existants: *Collister et Lahav 2004, M.Brescia et al. 2012,F.Feroz et al. 2014 ...* mais pas d'optimisation d'architecture de réseaux de neuronnes.

Solution Proposée par C. Arouri

3 étapes:

1. Application d'un algorithme de clustering non paramétrique: *X-Means, DBSCAN, Meanshift*.
2. Construction d'un RNA à 3 couches avec la couche cachée ayant autant de neurones que de groupes retournés par l'étape précedente.
3. Apprentissage du réseau : *algorithme MLPQNA.*

Estimation de Redshifts photométriques => Problème de regression. Des travaux existants: *Collister et Lahav 2004, M.Brescia et al. 2012,F.Feroz et al. 2014 ...* mais pas d'optimisation d'architecture de réseaux de neuronnes.

Solution Proposée par C. Arouri

3 étapes:

1. Application d'un algorithme de clustering non paramétrique: *X-Means, DBSCAN, Meanshift.*
2. Construction d'un RNA à 3 couches avec la couche cachée ayant autant de neurones que de groupes retournés par l'étape précedente.
3. Apprentissage du réseau : *algorithme MLPQNA.*

Solution appliquée à des données de tailles raisonnables.
Qu'en est-il des grandes masses données?

sommaire

Nous sommes à l'ère du "Big Data"

Big Data: Définition

Quatre dimensions(4 V):

1. **Volume:** Grandes quantités de données (ordre du zettaoctet: 10^{21} O).
2. **Variété:** Différents types de données provenant de différentes sources.
3. **Vélocité:** Frequence à laquelle les données sont générées, capturées et partagées.
4. **Véracité:** Plausibilité des informations générées.

Nous sommes à l'ère du "Big Data"

Big Data: Définition

Quatre dimensions(4 V):

1. **Volume:** Grandes quantités de données (ordre du zettaoctet: 10^{21} O).
2. **Variété:** Différents types de données provenant de différentes sources.
3. **Vélocité:** Frequence à laquelle les données sont générées, capturées et partagées.
4. **Véracité:** Plausibilité des informations générées.

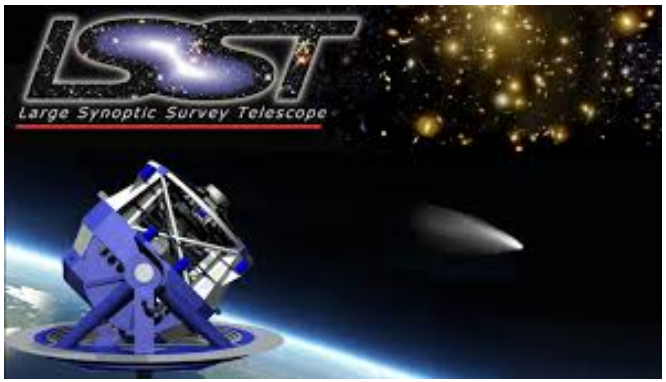
Exemple: nombre de galaxies

- 100 milliards: galaxies observables dans l'univers visible.
- Dizaines de millions déjà photographiées.

Projet PetaSky

Gestion et exploration des grandes masses de données scientifiques issues d'observations astronomiques grand champ.

1. **LSST** : 3,2 pixels de données par nuit pendant 10 ans => 140 Po d'images en fin de programme.



2. **Euclid:** mesurera les distances de 50 millions de galaxies par un spectographe infrarouge *(7 ans)*.

Etapes

Trois grandes étapes:

1. Etudier l'état de l'art sur les methodes de classification non supervisé et le passage à l'echelle, notamment à travers les platesformes comme *SPARK ou HadooP /MapReduce.*

Etapes

Trois grandes étapes:

1. Etudier l'état de l'art sur les methodes de classification non supervisé et le passage à l'echelle, notamment à travers les platesformes comme *SPARK ou HadooP /MapReduce.*
2. Etudier l'etat sur le deepLearning et notamment les problèmes d'optimisation d'architecture.

Etapes

Trois grandes étapes:

1. Etudier l'état de l'art sur les methodes de classification non supervisé et le passage à l'echelle, notamment à travers les platesformes comme *SPARK ou HadooP /MapReduce.*
2. Etudier l'etat sur le deepLearning et notamment les problèmes d'optimisation d'architecture.
3. Proposer une approche d'optimisation d'architecture neuronale.

Ce qui a déjà été fait

1. **Etape préliminaire:**
   - Comprendre le travail précédemment fait *(Mémoire de l'etudiante C. Arouri.)*
   - Reproduire les expérimentations (*Weka, ANN_Z,...*)

Ce qui a déjà été fait

1. **Etape préliminaire:**
   - Comprendre le travail précédemment fait *(Mémoire de l'etudiante C. Arouri.)*
   - Reproduire les expérimentations (*Weka, ANN_Z,...*)
2. **Etude des algorithmes proposés pour le clustering des Big Data**
   - *A survey of clustering Algorithms for Big Data: Taxonomy and Empirical Analysis, Fahad et al.*
   - *Big Data Clustering: A review, A. Seyed et al.*
   - *Scalable K-Means ++, Bahman Bahmani.*

Ce qui a déjà été fait

1. **Etape préliminaire:**
   - Comprendre le travail précédemment fait *(Mémoire de l'etudiante C. Arouri.)*
   - Reproduire les expérimentations (*Weka, ANN_Z,...*)
2. **Etude des algorithmes proposés pour le clustering des Big Data**
   - *A survey of clustering Algorithms for Big Data: Taxonomy and Empirical Analysis, Fahad et al.*
   - *Big Data Clustering: A review, A. Seyed et al.*
   - *Scalable K-Means ++, Bahman Bahmani.*
3. **Survey sur les plateformes de manipulation des Big Data**
   - *Hadoop MapReduce => Mahout, Oryx*
   - *Apache Spark => MLlib*
   - *Sparkling water = Spark + H20*

Ce qu'il reste à faire

- Etudier l'impact de la selection d'une partie de l'ensemble des données sur l'optimisation de l'architecture.

Ce qu'il reste à faire

- Etudier l'impact de la selection d'une partie de l'ensemble des données sur l'optimisation de l'architecture.
- DeepLearning: définition, fonctionnement,...

Ce qu'il reste à faire

- Etudier l'impact de la selection d'une partie de l'ensemble des données sur l'optimisation de l'architecture.
- DeepLearning: définition, fonctionnement,...
- Proposer et tester une approche d'optimisation d'architecture neuronale.