Classification automatique de sources astronomiques par cartes auto-organisatrices

Carole THIEBAUT[1] et Michel BOËR[1]

[1]CESR-OMP-CNRS,

9 av. du Colonel Roche, BP 4346, 31028 Toulouse Cedex 4

carole.thiebaut@cesr.fr, michel.boer@cesr.fr

Résumé – Cet article présente le développement d'un classifieur automatique basé sur les cartes auto-organisatrices de Kohonen pour les objets présents sur les images astronomiques. L'originalité de la méthode consiste à utiliser un classifieur non supervisé et à lui présenter directement les pixels constituant les objets à étudier, sans utiliser de paramètres décrivant ces objets. Nous présentons différentes normalisations possibles pour ces descripteurs originaux. Le classifieur ainsi construit est testé sur des images astronomiques simulées et réelles (images d'une base de données astronomiques et du télescope automatique TAROT). Pour les deux types d'images, la méthode est aussi performante que les méthodes supervisées. Pour les images TAROT, qui sont des images très bruitées, la définition d'un seuil de classification au-delà duquel l'efficacité du classifieur n'est plus acceptable est nécessaire.

Abstract – We present the development of an automatic classifier for sources detected on astronomical images based on the Kohonen's self-organizing maps. The originality of this work consists in utilizing an unsupervised method and in presenting directly to the classifier the pixels of the detected objects instead of characteristic parameters extracted from it. Different normalizations of the input vector are presented. The so-built classifier is tested on simulated and real images (images from an astronomical database and from a real instrument: the TAROT telescope). For both images types, the method is as effective as a supervised method. For TAROT images, which are very noised, we have to define a classification threshold under which the classification efficiency is no more reliable.

1. Introduction

La classification automatique des objets présents sur les images astronomiques consiste à attribuer à chaque objet une nature selon qu'il soit un objet ponctuel (étoile) ou un objet étendu (galaxie, objet saturé, objet diffus). Les relevés profonds des étoiles et des galaxies sur de larges zones du ciel permettent d'aborder d'une manière statistique les problèmes liés à la dynamique des structures galactiques, aux effets environnementaux sur la formation des galaxies et à la distribution à grande échelle de la matière dans l'univers. Avec l'arrivée des nouveaux instruments au sol et en vol dédiés au relevé spectroscopique, les catalogues ainsi créés sont plus complets et précis. La discrimination entre les étoiles et les galaxies sur les images de ces relevés est donc un enjeu essentiel pour les études extragalactiques. Ce problème est largement abordé en astronomie. Dès le début des années 90, l'intelligence artificielle a supplanté les approches bayésiennes [12] et paramétriques [10]. L'étude d'Odewahn [9] a montré que les perceptrons multicouches permettait de répondre à ce problème. Les classifieurs supervisés les plus utilisés aujourd'hui sont le perceptron multi-couches implanté dans le logiciel Sextractor [3] ou celui implanté dans NExT [1]. Notons également une large utilisation des arbres de décision intégrés dans des logiciels comme OC1 [11]. D'autres études basées sur des réseaux neuronaux non supervisés concurrencent ces précédents travaux [7, 8]. Dans cet article, nous présentons le développement d'un classifieur basé sur les cartes auto-organisatrices de Kohonen. La première partie est dédiée à la mise en place de ce classifieur. La seconde partie présente les résultats obtenus pour des images astronomiques simulées et réelles. Les conclusions sont exposés dans la dernière partie.

2. Mise en place du classifieur

2.1 Le guide de choix

Le classifieur que nous souhaitons développer devra s'intégrer dans une chaîne de traitement automatique d'images. Le cahier des charges associé est donc le suivant : rapidité de décision, pas d'intervention humaine dans le processus et possibilité de classer des objets indéfinis. Nous avons choisi d'utiliser une méthode non supervisée en suivant les travaux encourageants de Miller & Coe [8]. Même si l'interprétation des méthodes non supervisées n'est pas facile et directe, contrairement aux méthodes supervisées qui fournissent une probabilité d'appartenance aux différentes classes définies a priori, elles s'affranchissent de l'avis d'un expert humain et d'une définition préliminaire des différentes classes du problème. Elles permettent également d'étudier des objets indéfinis. Le choix des descripteurs est présenté en détails dans la section 2.3.

2.2 Les cartes auto-organisatrices

L'idée des cartes auto-organisatrices [6] est d'effectuer une cartographie de l'espace des données d'entrée de grande dimension n dans un espace régulier de neurones de faible dimension (1, 2 ou 3). Chaque neurone i de la carte est associé à un vecteur de référence appelé également vecteur poids de n-dimensions. L'ensemble des vecteurs de référence

forme un dictionnaire. Les neurones de la carte sont connectés aux neurones adjacents par une relation de voisinage, qui détermine la topologie ou la structure de la carte. Dans l'algorithme d'apprentissage SOM classique, la topologie et le nombre de neurones restent constants tout au long du processus d'apprentissage.

Le processus d'apprentissage est le suivant :
Chaque neurone j est associé à un vecteur poids W_j de telle sorte que la composante W_{ij} connecte le neurone j à la $i^{éme}$ composante du vecteur d'entrée x.

1) Un vecteur échantillon x de la base d'apprentissage défini par ses composantes I_i, $i \in [1, n]$ est présenté à la carte et on calcule la valeur D_j pour chaque neurone :
$$D_j = \sum_{i=1}^{n} \| I_i - W_{ij} \|$$
avec n la dimension de l'espace d'entrée, et $\|.\|$ la distance Euclidienne. On mesure ainsi la similarité du vecteur échantillon avec les vecteurs du dictionnaire. Le neurone c associé la plus petite valeur D_j est le neurone gagnant. Ce neurone est donc celui dont les poids sont les plus proches des composantes du vecteur d'entrée.
2) Après avoir sélectionné le neurone gagnant, on met à jour les vecteurs poids de ses neurones voisins selon l'équation : $W_j(t+1) = W_j(t) + h_{ci}(t) \times (x - W_j(t))$ avec t le temps, et h_{ci} la fonction de voisinage associée au neurone c.
3) Les étapes 1 et 2 sont répétées jusqu'à la fin de l'apprentissage. Le nombre d'époques de la phase d'apprentissage est défini à l'avance.

La fonction de voisinage utilisée pour notre étude est le noyau gaussien proposé par Kohonen [6]. Une initialisation linéaire des vecteurs poids est adoptée : ils sont choisis dans l'espace vectoriel à deux dimensions engendré par les vecteurs propres de la matrice d'autocorrélation des données correspondant à ses 2 valeurs propres les plus élevées. La carte étant approximativement organisée dès le début, cette initialisation accélère la phase d'apprentissage. Nous choisissons une carte bidimensionnelle de forme hexagonale pour une meilleure visualisation. Le nombre de neurones de la carte dépend du nombre N de vecteurs d'apprentissage (il est égal à $5\sqrt{N}$).

Lorsque l'apprentissage est terminé, chaque vecteur de la base de test est présenté à ce classifieur qui identifie le neurone gagnant associé en calculant la distance entre chaque vecteur poids de tous les neurones de la carte et le vecteur des descripteurs de l'objet présenté en entrée. L'objet fera donc partie de la région du neurone gagnant identifié. Une analyse visuelle de la carte ou un algorithme de regroupement a posteriori est donc nécessaire à ce stade pour identifier la nature de chaque objet de la base de test. Dans notre étude, nous labellisons les objets de la base d'apprentissage en identifiant chaque objet avec un catalogue astronomique. Chaque neurone de la carte sera donc labellisé, après apprentissage par le label du nombre maximum d'objets ayant fait réagir le neurone.

2.3 Les descripteurs

Les descripteurs utilisés sont les pixels constituant l'objet étudié. Ce mode de présentation directe a déjà été utilisé dans [7] et [2]. La taille des images est 30×30 pixels pour la première étude et 17×17 pixels pour la seconde. Pour l'étude des images TAROT [4, 5], la taille des sous-images utilisées dépend de l'échantillonnage et de la réponse instrumentale. La résolution des images TAROT est 3.6 arcsec et les images sont sur-échantillonnées de sorte que les objets occupent en moyenne 9 pixels. Le nombre de descripteurs doit être le plus faible possible pour que l'algorithme d'apprentissage soit rapide. La taille de la sous-image choisie pour notre étude est donc de 11×11 pixels, i.e. juste un peu plus grande que la surface moyenne occupée par les sources observées par l'instrument TAROT. Les descripteurs de notre étude sont donc 121 pixels ; nous parlerons indifféremment de sous-image d'entrée de 11×11 pixels ou de vecteur d'entrée de 121 pixels.

Les images TAROT étant codées sur 16 bits non signés, les pixels peuvent prendre une valeur comprise entre 0 et 65535. Le vecteur des descripteurs correspondant à une sous-image doit donc être normalisé de façon à ne pas présenter de grandes variations et donc à ne pas biaiser l'algorithme d'apprentissage qui favorise les grandes valeurs de descripteurs. Nous utilisons trois types de normalisation : celle utilisée dans [7] qui transforme le vecteur à moyenne nulle et variance unité (appelée *Norm1* pour notre étude), les normalisations « average » et « grid » mises au point dans [2] (appelées *Norm2* et *Norm3* pour notre étude). La normalisation « average » n'est pas « linéaire » car, contrairement aux deux autres, elle fait correspondre aux 121 pixels un vecteur de 21 composantes.

3. Les résultats

Nous testons le classifieur ainsi construit sur des images astronomiques simulées et réelles. Les résultats sont obtenus en utilisant la boîte à outils Matlab « SOM Toolbox » (*http://www.cis.hut.fi/projects/somtoolbox/*) développée par l'université d'Helsinki.

3.1 Sur les images simulées

Les images sont simulées avec le logiciel Skymaker développé par E. Bertin (*http://terapix.iap.fr/soft/skymaker/*). Nous construisons deux types d'images : des images avec le système optique réglé au foyer et une deuxième série d'images avec système déréglé. Les objets présents sur les deux types d'images ont des caractéristiques différentes : pour une optique mal réglée, les objets occupent plus de pixels (réponse instrumentale plus étalée). Sur chaque image, nous simulons des profils d'objet de type galaxie et étoiles.

Les bases de test et d'apprentissage sont constituées d'objets extraits des images simulées. *L'étude 1* correspond aux images avec système optique déréglé : 1316 objets dans la base d'apprentissage (1258 S + 58 G), 663 objets (634 S + 29 G) de test. *L'étude 2* correspond aux images avec système optique au foyer : 1767 objets (1623 S + 144 G) dans la base

d'apprentissage, 830 objets (772 S + 58 G) de test. *L'étude 3* correspond aux deux types d'images.

Pour illustrer les résultats, nous présentons, Figure 1, la U-matrice (carte des distances entre les unités voisines) obtenue pour la *Norm1* et l'étude 1 et cette carte obtenue après un algorithme de regroupement des k-moyennes (« k-means clustering ») en 10 régions. Nous observons différentes régions sur cette carte délimitée par des fortes valeurs de distances. La *Norm1* utilise des vecteurs de descripteurs de 121 composantes. Nous pouvons donc visualiser les poids des neurones sous la forme de sous-images de 11×11 pixels (voir figure 2). Nous observons alors les objets relatifs aux différentes régions mises en évidence par la U-matrice.

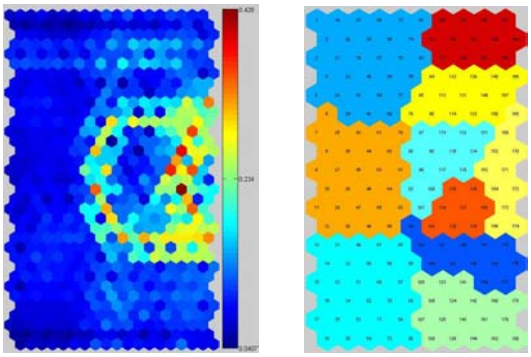

Fig.1 : Gauche - U-Matrice obtenue pour l'étude 1 et la *Norm1*. Droite - U-Matrice après un algorithme des 10-moyennes sur les unités de cette carte.

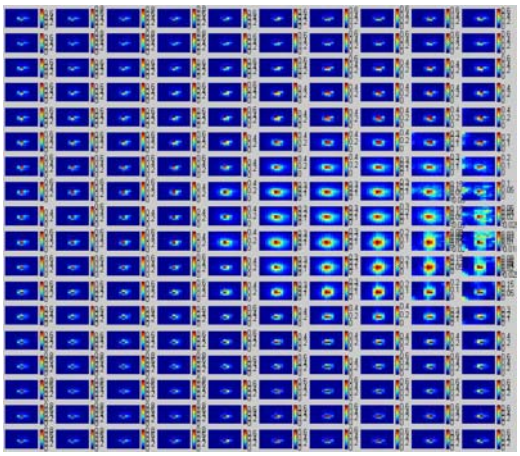

Fig. 2 : Cartes des poids des neurones visualisés comme des sous-images de taille 11×11 pixels.

Pour quantifier l'efficacité de la classification et la comparer à celle du perceptron implanté dans SExtractor, nous utilisons les notions introduites dans [2] : la ***sensibilité*** est le nombre total de galaxies correctement identifiées sur le nombre de galaxies présentes dans la base de test, la ***spécificité*** est le même rapport appliqué aux étoiles, la ***valeur de prédiction positive*** (PPV) est le nombre de galaxies correctement identifiées sur le nombre total de galaxies identifiées et la ***valeur de prédiction négative*** (NPV) est le même rapport appliqué aux étoiles. La table 1 présente les résultats obtenus pour les deux séries d'images et une série constituée de toutes les images simulées.

TAB. 1 : Tableau des résultats pour les images simulées avec système optique réglé et déréglé.

	Sensibilité	Spécificité	PPV	NPV
Etude 1				
Norm1	93.10%	99.84%	96.43%	99.68%
Norm2	51.72%	100%	100%	97.84%
SExtractor	*96.56%*	*46.06%*	*7.57%*	*99.66%*
Etude 2				
Norm1	96.55%	99.74%	96.55%	99.74%
Norm2	75.86%	97.93%	73.33%	98.18%
SExtractor	*81.03%*	*79.01%*	*22.48%*	*98.22%*
Etude 3				
Norm1 (déréglé)	79.31%	98.90%	76.67%	99.05%
Norm1 (réglé)	96.55%	99.74%	96.66%	99.74%
Norm2 (déréglé)	65.52%	100%	100%	98.45%
Norm2 (réglé)	0%	100%	0%	93.23%
SExtractor(déréglé)	*96.56%*	*46.06%*	*7.57%*	*99.66%*
SExtractor (réglé)	*81.03%*	*79.01%*	*22.48%*	*98.22%*

Pour l'étude 1, la *Norm1* présente de très bons résultats pour les 4 valeurs d'efficacité. La *Norm2* présente une faible sensibilité aux galaxies mais les 3 autres valeurs sont très bonnes. Les résultats du classifieur de SExtractor présentent une très faible valeur de prédiction positive i.e. que le nombre de galaxies identifiées au total est élevé, le classifieur classe donc beaucoup d'étoiles en galaxies, ce qui est logique puisque le système est défocalisé donc les objets déformés et allongés. Cette thèse est confirmée par la faible valeur de spécificité (peu d'étoiles bien identifiées).

Pour l'étude 2, les résultats de la *Norm1* sont meilleurs que ceux de la *Norm2* (98.14% en moyenne contre 86.32%). Les résultats de SExtractor présentent trois bonnes valeurs mais une faible valeur de prédiction positive.

Pour l'étude 3, les résultats de la *Norm1* sont très bons et meilleurs que ceux de SExtractor, par contre, les résultats obtenus avec la *Norm2* sont très critiques avec 0% de sensibilité et de valeur de prédiction positive.

3.2 Sur les images réelles du DSS

Le « Digitized Sky Survey » (par la suite DSS) est une base de données d'images optiques du ciel de taille 1°×1° acquises lors du relevé photographique du télescope Schmidt et numérisées par la suite. Ce catalogue est accessible sur le site Web du « Space Science Telescope Institute » : *http://skyview.gsfc.nasa.gov*. Nous téléchargeons les images des objets indiqués dans [2]. Ces objets ont été identifiés par Steve Odewahn qui a confirmé l'identité des 60 galaxies et 27 étoiles étudiées. Pour cette étude, le nombre de données étant restreint, nous utilisons la méthode du « leave-one-out » : la base de test est constituée d'un seul objet et les 86 autres représentent la base d'apprentissage. Nous comparons les résultats de notre classifieur (*Norm1* et *Norm3*) avec les

algorithmes de « Learning Vector Quantization » et de rétro-propagation déjà testés sur ces images [2].

TAB. 2 : Comparaison de l'efficacité de notre classifieur avec celle des algorithmes « Learning Vector Quantization » et rétro-propagation.

	SOM *Norm1*	SOM *Norm3*	LVQ	Rétro-Propagation
Sensibilité	90%	88.3%	87%	97%
Spécificité	92.6%	96.3%	96%	96%
PPV	96.4%	98.1%	98%	98%
NPV	80.6%	78.8%	76%	93%

Les valeurs des quatre paramètres obtenues pour notre classifieur topologique sont satisfaisantes. Les deux normes utilisées pour cette étude présentent des résultats similaires avec des valeurs moyennées sur les quatre paramètres d'efficacité très proches (89.9% pour la *Norm1* et 90.37% pour la *Norm3*). L'efficacité de notre réseau, quelque soit la norme utilisée, est comparable à celle de l'algorithme « Learning Vector Quantization » (89.25% de succès moyenné pour les quatre paramètres). Ceci vérifie l'hypothèse selon laquelle cet algorithme est une des méthodes supervisées les moins efficaces. Par contre, l'algorithme de rétro-propagation est le plus efficace des trois algorithmes étudiés ici (96% de succès moyenné pour les quatre paramètres). Cette méthode nécessite cependant une expertise préalable dont nous voulons nous affranchir.

3.3 Sur les images TAROT

L'étude des images du télescope TAROT est en cours de développement. Dans un premier temps, nous utilisons notre classifieur comme un algorithme de regroupement sur une base d'apprentissage constituée d'objets TAROT. Les images sélectionnées pour cette étude doivent présenter une densité d'objets de type galaxie supérieure à la moyenne (sinon, le nombre de galaxies est très faible par rapport au nombre d'étoiles). Nous utilisons donc des zones d'amas de galaxies. Nous analysons cette base d'apprentissage avec notre réseau topologique. A l'aide du catalogue astronomique Simbad (*http://simbad.u-strasbg.fr*), nous déterminons la nature des objets de type galaxie présents sur l'image et nous labellisons ces objets. Lorsqu'on visualise les objets ayant fait réagir les neurones labellisés « galaxie », nous observons une grande population d'objets de type « étoiles ». En fait, les objets faiblement lumineux de nos images sont très proches du fond de ciel et donc entaché du bruit de nos images qui est assez important (le bruit de lecture de l'instrument TAROT est de 14e-). Ceci explique les mauvais résultats de la classification de SExtractor qui trouvait pour chaque image entre 75% et 90% d'objets de type galaxie. Nous devons donc tenir compte de la qualité de l'image étudiée et déterminer un seuil de classification critique, lié à cette qualité d'image. Au-delà de ce seuil, l'efficacité de la classification n'est plus acceptable. Nous utilisons la magnitude instrumentale de l'objet détecté sur l'image.

4. Conclusions

Nous avons présenté le développement d'un classifieur automatique pour les sources présentes sur les images astronomiques. Le classifieur est développé pour les images TAROT mais peut être utilisé pour les images d'autres instruments. Si l'échantillonnage angulaire des autres instruments est comparable à celui de TAROT, aucune modification n'est nécessaire. Dans le cas contraire, la taille de la sous-image centrée sur l'objet étudié dépendra des caractéristiques du nouvel instrument.

La méthode est non supervisée et utilise une carte auto-organisatrice de Kohonen. Les descripteurs associés sont les pixels constituant l'objet étudié (sous-image de 11×11 pixels) normalisés de trois façons différentes : deux normalisations dites linéaires qui associent à ces pixels un vecteur de 121 composantes et une normalisation non linéaire qui associe un vecteur de 21 composantes. Nous avons présenté l'efficacité de notre classifieur pour des images simulées et des images réelles issues d'une base de données d'images astronomiques. Les résultats obtenus sont comparables au perceptron multicouches implanté dans SExtractor (pour les images simulées) et aux algorithmes supervisés (pour les images réelles). L'étude préliminaire des images TAROT nous montre que le niveau de bruit assez élevé de ces images pose problème à tout type de classifieur, les objets ponctuels proches du bruit de fond de l'image étant souvent considérés comme des objets étendus. Pour ces images, nous devons donc définir un seuil de classification optimal au-delà duquel l'efficacité de la classification ne sera pas considérée comme satisfaisante. Nous envisageons d'utiliser la magnitude instrumentale de l'objet détecté sur l'image. Cette valeur contient l'information de la qualité de l'image et nous permettra d'étudier l'évolution de l'efficacité de la classification en fonction de la dégradation de l'image.

Références

[1] Andreon S. et al., *Mon. Not. R. Astron. Soc.*, 319, 700-716, 2000.

[2] Bazell D. & Peng Y., *Astrophys. Journ. Supp. Ser.,* 47-55, 1998.

[3] Bertin E. & Arnouts S., *Astron. & Astrophys. Supp. Ser.,117, 393-404*, 1996.

[4] Boër M. et al., *Astron. & Astrophys. Supp. Ser.*, 138, 579-580, 1999.

[5] Bringer M. et al., *Experim. Astron.*,12, 1, 33-48, 2001.

[6] Kohonen. T, Self-Organizing Maps, 2nd edition, *Springer Series in Information Sciences*, 1997.

[7] Mähönen P.H. & Hakala P.J., *Astrophys. Journ.,* 452, L77-L80, 1995.

[8] Miller A.S. & Coe M.J., *Mon. Not. R. Astron. Soc.*, 279, 293-300, 1996.

[9] Odewahn S., et al., *Astron. Joun.*, 103, 1, 318-331, 1992.

[10] Reid N. & Gilmore G., *Mon. Not. R. Astron. Soc.*, 201, 73, 1982.

[11] Salzberg S. et al., *Astron. Soc. of the Pacific*, 107, 279-288, 1995.

[12] Sebok W.L., *The Astronomical Journal*, 84, 1526, 1979.