Chapitre 9

Les résultats d'apprentissages, quelques visualisations

Dans ce chapitre, nous décrivons un exemple de classification afin, d'une part, de vérifier les propriétés que nous prévoyons dans la partie 7.1 et d'autre part, de mettre en avant les différences inévitables que présentent deux classifications issues d'un apprentissage similaire[1], dues aux processus aléatoires qui y interviennent (voir l'algorithme d'apprentissage des SOM, dans la section 3.4.3.1). Nous insistons donc sur le fait que les classifications que nous obtenons n'ont rien d'absolu.

9.1 Détail d'une classification : les articles de A&A, de 1994 à 1999

Afin d'illustrer le type de classifications que permet l'utilisation des SOM sur des documents bibliographiques, nous allons décrire celle que nous avons obtenue avec les articles parus dans la revue de référence européenne Astronomy and Astrophysics (A&A). Il s'agit de l'une des classifications accessibles[2] au CDS pour laquelle l'apprentissage a été réalisé avec les articles parus entre 1994 et 1996, décrits par 269 mots-clés. Depuis, cette carte est mise à jour par ajout simple des nouveaux documents (voir le chapitre 8).

9.1.1 La visualisation de thèmes généraux

Nous avons choisi d'étudier la localisation des documents relatifs aux thèmes suivants : la cosmologie, les galaxies, le milieu interstellaire, les étoiles et le soleil. Ces cinq grands domaines de l'astrophysique sont fréquemment abordés dans les publications de A&A. Afin de visualiser les zones concernées, nous avons effectué une requête par mots-clés, en recherchant pour chacun de ces thèmes, les documents qui contiennent au moins un des dix mots-clés les plus fréquents en rapport avec ce thème (tableau 9.1).

La figure 9.4 montre la localisation des documents tels que nous l es avons sélectionnés pour chaque thème. Sur ces images, le code de couleur indique la densité des documents :

[1] Nous appelons apprentissages similaires les apprentissages utilisant les mêmes paramètres et les mêmes données, qui diffèrent uniquement par les mécanismes aléatoires qu'ils mettent en jeu.

[2] L'adresse officielle est http ://simbad.u-strasbg.fr/A+A/map.pl, mais une version plus récente de l'interface est accessible à l'adresse http ://newb6.u-strasbg.fr/ poincot/cgi-bin/Cartes/bib.pl, la plus ancienne ne dispose pas des carrés de taille variable, ni du système de requêtes par bibcode.

le soleil	les étoiles	le milieu interstellaire
sun : magnetic fields	stars : circumstellar matter	ISM : molecules
sun : corona	X-rays : stars	ISM : dust,extinction
sun : oscillations	stars : abundances	ISM : clouds
sun : flares	stars : evolution	radio lines : ISM
sun : chromosphere	stars : mass loss	ISM : jets and outflows
sun : solar wind	stars : binaries : close	Infrared : ISM : lines and bands
sun : activity	stars : pre-main sequence	ISM : abundances
sun : sunspots	stars : AGB and post-AGB	ISM : planetary nebulae : general
sun : radio radiation	stars : fundamental parameters	ISM : kinematics and dynamics
sun : photosphere	stars : atmospheres	ISM : HII regions
les galaxies	**la cosmologie**	
galaxies : ISM	cosmology : dark matter	
galaxies : active	cosmology : gravitational lensing	
galaxies : kinematics and dynamics	cosmology : large scale structure of Universe	
radio continuum : galaxies	cosmology : observations	
galaxies : evolution	cosmology : theory	
X-rays : galaxies	cosmology : cosmic microwave background	
galaxies : Magellanic Clouds	cosmology : diffuse radiation	
galaxies : clusters	cosmology : distance scale	
galaxies : jets	cosmology : miscellaneous	
galaxies : nuclei	cosmology	

Tab. 9.1: *Les 10 plus fréquents mots-clés relatifs aux thèmes choisis.*

- le fond bleu uni, signifie qu'aucun document n'est sélectionné ;
- les couleurs dégradées allant du bleu-vert, jusqu'au rouge[3] signifient un nombre de documents croissant.

Nous nous intéressons également aux carrés qui correspondent à chaque classe. Leur taille est reliée au nombre maximal de mots-clés de la requête, contenus simultanément dans un même document.

Remarque : Sur chaque image, la zone qui contient le plus de documents est de couleur rouge[3]. Les couleurs de ces cartes ne permettent pas de comparer quantitativement les groupes visibles sur deux cartes différentes, car elles sont relatives à la densité maximale d'une carte donnée, variable d'une image à l'autre.

9.1.2 La classification globale

De manière générale, on remarque que les thèmes que nous avons visualisés se retrouvent sur des zones différentes de la carte, couvrant ainsi la majeure partie de sa surface. La plupart des documents relatifs à un thème sont regroupés dans des zones denses et bien délimitées. Généralement, ils contiennent une fraction importante des mots-clés utilisés dans la requête (les carrés sont de taille importante dans ces régions). Notons que les zones relatives aux galaxies et au milieu interstellaire semblent divisées respectivement en trois et deux groupes. Une reconfiguration (voir 7.2.4) appropriée de la carte montrerait ces zones sous une forme non dissociée.

Autour de ces zones denses sont dispersés des documents, moins denses, qui comportent globalement moins de mots-clés relatifs au thème visualisé (carrés de petite taille). Ces documents ont donc certainement un rapport plus lointain avec le thème choisi. Leur quantité dépend assurément du nombre de mots-clés utilisés pour nos requêtes. Ainsi, si nous reproduisons l'expérience avec davantage de mots-clés, nous assistons à : a) un accroissement limité du nombre de ces documents, dispersés, peu denses et relativement peu en rapport avec le thème choisi ; b) à un accroissement du nombre de mots-clés contenus simultanément dans les documents des zones denses (Fig. 9.1).

[3]Il est possible que les couleurs n'apparaissent pas dans cet ouvrage de la même manière qu'à l'écran, on se reportera à la figure 7.4 qui montre comment la densité des documents est reliée aux couleurs.

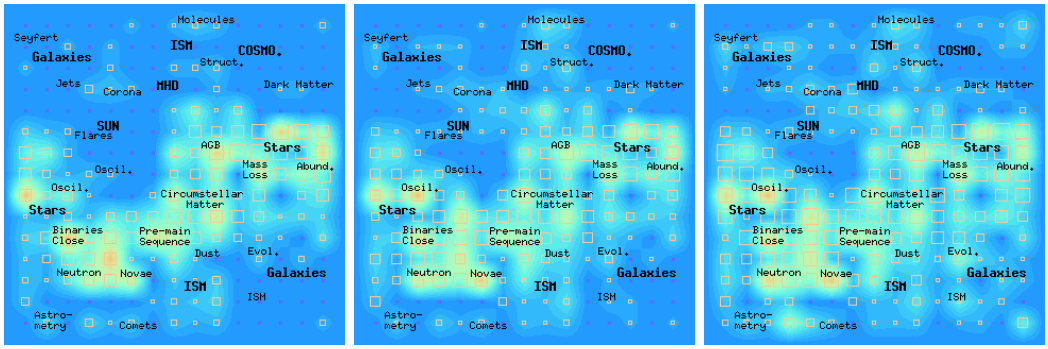

Fig. 9.1: *Influence du nombre de mots-clés (10, 20 et 40) utilisés dans les requêtes pour la visualisation des thèmes.*

9.1.3 La taille des zones

S'il est impossible de comparer la quantité de documents dans les zones visibles sur différentes images, on peut tout de même comparer la surface qu'elles occupent.

Nous constatons que les zones visualisées sur la figure 9.4 sont plus ou moins étendues suivant le thème abordé. Nous y observons en effet que :

- la cosmologie et le soleil sont localisés sur des régions peu étendues ;
- les galaxies et le milieu interstellaire occupent davantage de surface ;
- les étoiles forment le thème le plus étendu.

La taille des régions relatives à chaque thème reflète la diversité et le nombre de documents dans lequel il est abordé. Le tableau 9.2 montre en effet que parmi les mots-clés les plus fréquents, ceux relatifs aux étoiles sont les plus nombreux, suivis par les galaxies, le milieu interstellaire, le soleil et la cosmologie. Les sujets qui s'étendent sur de larges régions sont donc ceux sur lesquels il existe un grand nombre de documents décrits par un grand nombre de mots-clés.

sujet	sur les 30 mots-clés les plus fréquents	sur les 100 mots-clés les plus fréquents
la cosmologie	0	4
les galaxies	4	20
le milieu interstellaire	5	13
les étoiles	16	38
le soleil	1	6

Tab. 9.2: *Nombre de mots-clés relatifs aux différents thèmes sélectionnés, parmi les 30 et les 100 mots-clés les plus fréquents. Les mots-clés manquant sont ceux qui ne s'inscrivent pas exclusivement dans l'un de ces thèmes tels que "MHD", "polarization" ou encore "methods : numerical".*

9.1.4 Les recouvrements

La figure 9.4 permet d'observer les recouvrements de certains domaines, que constituent les classes qui possèdent des documents relatifs soit à deux thèmes à la fois, soit à l'un ou à l'autre. D'une manière générale, les deux thèmes sont abordés dans les mêmes documents, comme nous allons le voir dans la suite.

Les étoiles et le soleil

Les thèmes des étoiles et du soleil (images d et e) partagent une petite zone, qui fait le lien entre les régions qu'ils occupent. On y trouve des articles sur la constitution des étoiles et du soleil ("stars :interiors", "stars : atmosphere", "sun :atmosphere"), et les ondes sismiques qui s'y développent ("stars :oscillations", "sun :oscillations", "hydrodynamics").

Les galaxies et la cosmologie

Le thème de la cosmologie apparaît très lié à celui des galaxies. Les images a et b montrent même la cosmologie comme un sujet englobé dans le thème des galaxies. Ceci s'explique par le lien étroit qui relie ces deux thèmes. La cosmologie est en effet l'étude des grandes structures dans l'Univers ("galaxies : clusters"), de sa formation ("galaxies : formation"). Un sujet de recherche important en cosmologie est aussi celui de la recherche de la matière noire dans le halo de la Galaxie par l'étude du mouvement des étoiles ("galaxy :halo", "galaxies :kinematics and dynamics").

Le milieu interstellaire et les galaxies

Les thème du milieu interstellaire et celui des galaxies (images b et c) partagent des documents. Il y est question d'observations de molécules présentes dans le milieu interstellaire de différentes galaxies ("ISM :molecules", "galaxies :ISM"). L'évolution des galaxies est également abordée, comme la formation d'étoiles ("galaxies :starsburst", "ISM :clouds", "ISM :molecules").

Le milieu interstellaire et les étoiles

Ces deux thèmes (images c et d) sont liés principalement par des articles qui traitent des régions HII qui apparaissent autour des étoiles en formation ("stars :formation", "ISM :HII regions"), et par l'influence des éjections des étoiles évoluées, jusqu'aux nébuleuses planétaires, sur la composition du milieu interstellaire ("stars :mass loss","stars :AGB and post-AGB", "ISM :planetary nebulae :general","ISM :dust,extinction").

Les étoiles et les galaxies

Ces deux thèmes partagent une zone (images d et b) qui contient principalement des articles sur l'étude d'étoiles dans les galaxies voisines des Nuages de Magellan ("galaxies :Magellanic Clouds"). Ces études vont de la détermination des propriétés d'étoiles individuelles ("stars :fundamental parameters", "stars :luminosity function,mass function") à celle d'amas stellaires ("galaxies :stars clusters").

9.1.5 L'organisation des documents d'un thème, l'exemple des étoiles

Afin de montrer les caractéristiques locales de la classification, nous allons nous intéresser à la répartition des documents relatifs à un thème. Nous avons choisi pour cela le thème des étoiles, dont nous allons décrire le contenu de certaines classes remarquables, réparties le long d'un parcours représenté sur la figure 9.2. Voici les sujets abordés dans ces classes :

- les paramètres fondamentaux des étoiles, notamment la détermination de la masse des étoiles de systèmes binaires (à éclipses ou spectroscopiques) ;

- les oscillations dans les binaires spectroscopiques, puis les oscillations stellaires en général ;
- les binaires proches ;
- les binaires proches dont une composante est une étoile à neutrons, puis les étoiles à neutrons en général ;
- les étoiles entourées d'un disque d'accrétion, émettant des rayons X ;
- les variables cataclysmiques et leur disque d'accrétion, et également les émissions en rayons X ;
- les étoiles pré séquence principale, leur émission X, la formation d'étoiles ;
- les étoiles à émission, et la matière circumstellaire ;
- les étoiles AGB et post-AGB, la matière circumstellaire ;
- la perte de masse des étoiles, les étoiles super-géantes et des étoiles précoces ;
- les abondances dans les étoiles précoces, puis les abondances stellaires en général ;
- finalement, les abondances et les paramètres fondamentaux stellaires sont abordés sur le bord de la carte, ce qui nous ramène par continuité sur le bord de la carte à notre point de départ : les paramètres fondamentaux stellaires.

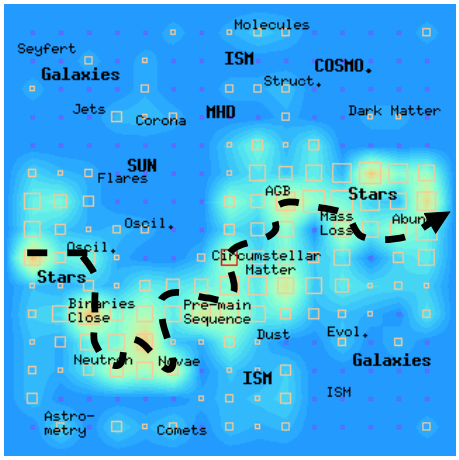

Fig. 9.2: *Parcours sur la carte, de la description des classes de documents relatifs aux étoiles.*

Remarque : les cartes détaillées proposent le même type de classification. Par exemple, la carte détaillée qui correspond à la zone du départ de la description que nous venons de faire, permet la séparation des articles relatifs aux binaires à éclipses, de ceux qui traitent des binaires spectroscopiques (Fig 9.3).

9.2 Le point de convergence des apprentissages : comparaison de deux classifications

Nous avons vu que l'algorithme d'apprentissage des SOM fait fait appel à des processus aléatoires (voir 3.4.3.1). Il en résulte que deux apprentissages pratiqués sur les mêmes données n'aboutiront pas exactement la même classification.

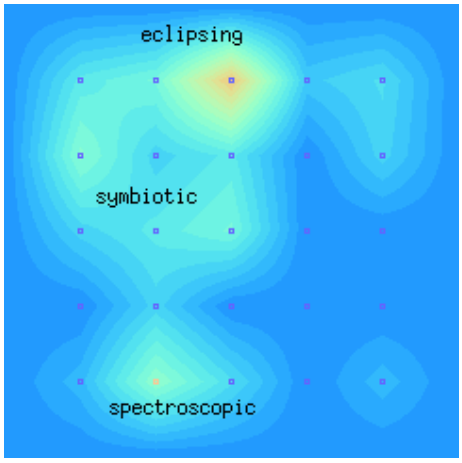

Fig. 9.3: *Carte détaillée des articles relatifs aux paramètres fondamentaux de binaires à éclipses ou spectroscopiques.*

Nous avons effectué deux apprentissages sur les données constituées par les articles parus dans A&A entre 1994 et 1999, décrits par 269 mots-clés. Les résultats obtenus vont nous servir à étudier les différences constatées sur l'organisation générale des grands thèmes sur la carte, et sur les classes de documents résultantes.

9.2.1 L'organisation globale

La figure 9.5 montre les documents relatifs aux cinq thèmes dont il a été question dans la partie 9.1.1. Les documents visualisés sont les mêmes pour les deux classifications. On remarque que :

- les positions relatives des divers thèmes sont conservées d'une classification à l'autre ;
- la taille des régions occupées par les mêmes thèmes dans les deux classifications sont comparables.

Globalement, les classifications obtenues sont donc assez semblables. Nous pouvons même constater qu'elles sont également assez similaires à la classification que nous avons décrite dans la section 9.1.1, qui résulte d'un apprentissage effectué sur environ deux fois moins de données (les articles publiés de 1994 à 1996). On remarquera par exemple les dimensions comparables des zones où apparaissent les différents thèmes, ou encore le lien très étroit entre la cosmologie et les galaxies.

9.2.2 Les classes de documents

La figure 9.6 montre la localisation de documents dans les deux classifications que nous comparons. Nous avons choisi les documents de certaines classes de la première classification (images de la première ligne), pour les visualiser dans la seconde classification (images de la seconde ligne).

9.2.2.1 La dispersion des classes

Nous remarquons que les documents regroupés dans une même classe dans la première classification ne le sont pas forcément dans la seconde.

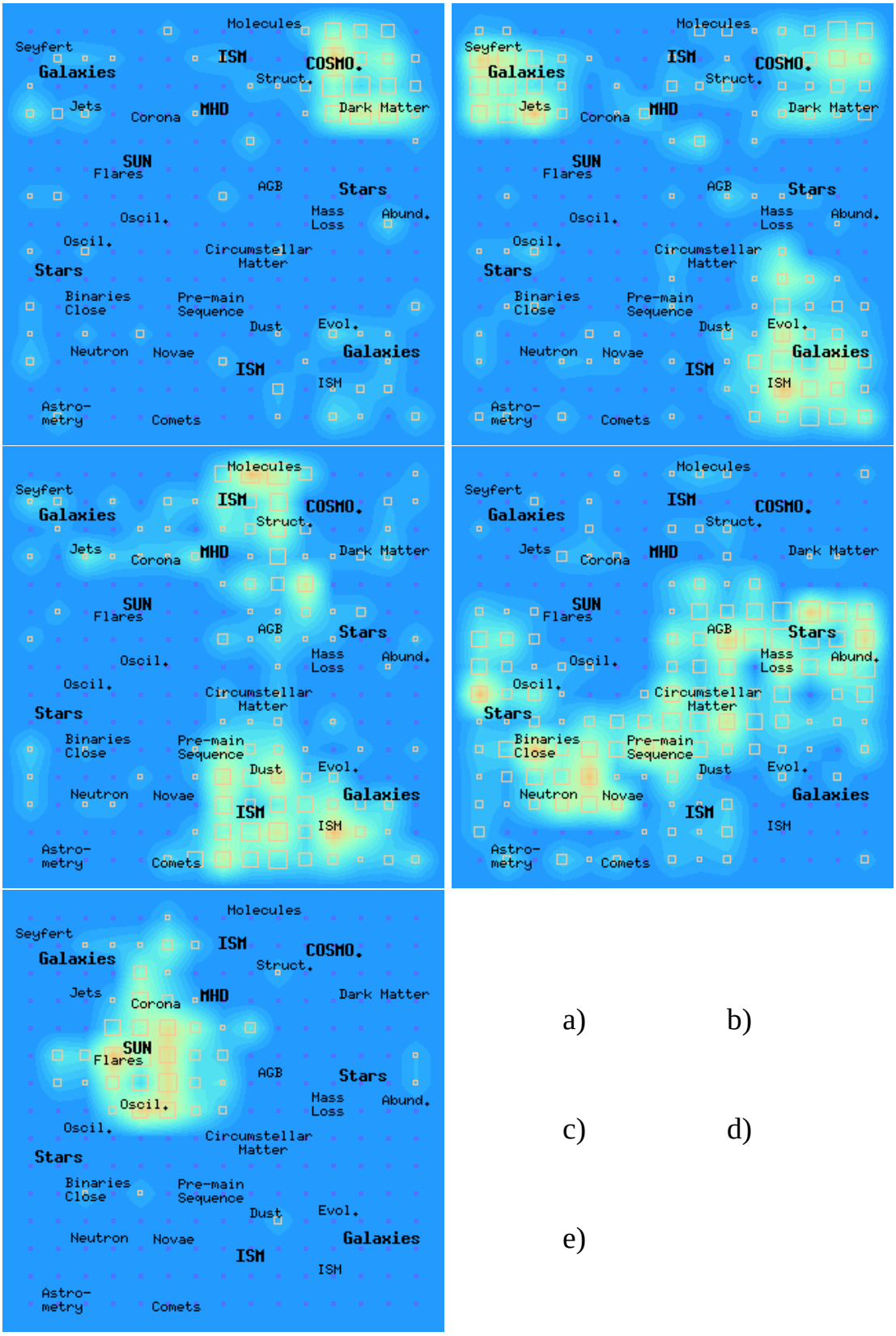

Fig. 9.4: *Localisation des documents contenant les 10 mots-clés les plus fréquents relatifs à : a) la cosmologie, b) les galaxies, c) le milieu interstellaire, d) les étoiles e) le soleil.*

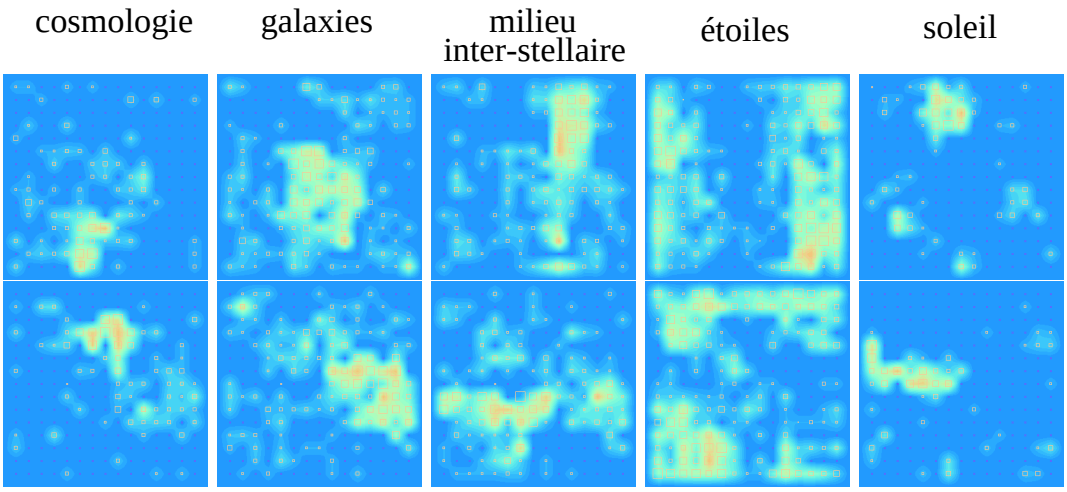

Fig. 9.5: *Localisation des thèmes (suivant le même procédé qu'en 9.1.1) pour deux classifications résultant d'apprentissages similaires. Les images situées sur la même ligne correspondent à une même classification.*

Généralement, on observe que la majeure partie des documents se retrouvent dans la même classe, tandis que d'autres se retrouvent dans le voisinage des premiers. Cela montre une distorsion des classes correspondantes dans les deux classifications. Ces classes ne sont pas tout à fait identiques.

D'autres documents sont localisés plus loin encore du groupe principal. Ceux-ci sont décrits par des mots-clés relatifs à plusieurs thèmes d'importance similaire. Selon la classification, ils se retrouvent dans une classe relative à l'un ou l'autre de ces thèmes(Tab. 9.3).

description de l'article	principaux mots-clés de la classe 1	principaux mots-clés de la classe 2
stars : late type	molecular processes	stars : late type
stars : atmospheres	ISM : molecules	stars : atmospheres
stars : low mass, brown dwarfs	ISM : clouds	stars : fundamental parameters
molecular processes	ISM : abundances	
Infrared : stars	ISM : dust,extinction	
molecular data	molecular data	

Tab. 9.3: *Les mots-clés principaux des classes auxquelles appartient l'article ([4]) dans les deux classifications.*

Nous avons évalué à près de 70% les documents qui appartiennent à la même classe dans les deux classifications. Si on ajoute à ces documents, ceux qui se trouvent dans une classe adjacente, ce nombre est porté à 80% environ. Les documents 20% de documents restant sont ceux qui se trouvent dans des classes très différentes (et éloignées) dans les deux classifications, comme l'article du tableau 9.3.

[4]Model atmospheres of cool, low-metallicity stars : the importance of collision-induced absorption. (BORYSOW A., JORGENSEN U.G., ZHENG C.)

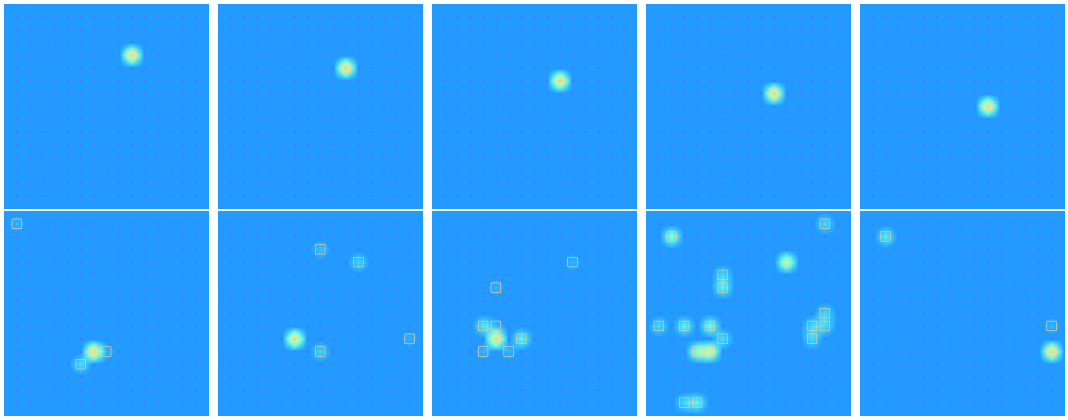

Fig. 9.6: *Localisation des documents des classes choisies dans la première classification (première ligne) et leur localisation dans la seconde classification (seconde ligne).*

9.2.2.2 La localisation des classes

Les cinq classes que nous avons choisies sont des classes voisines, alignées suivant la même colonne (Fig. 9.6). Considérons la localisation des documents de ces classes dans le seconde classification :

- les trois premiers groupes de documents se situent principalement dans trois classes voisines, avec une faible dispersion des documents ;
- le quatrième groupe de documents se trouve très dispersé autour d'une classe voisine des trois précédentes Dans la première classification, ce groupe fait typiquement partie d'une classe située entre deux thèmes assez éloignés. Cette classe est elle-même assez peu homogène, puisque sur les 42 documents qu'elle contient, un mot-clé apparaît dans 41 documents (ISM : Bubbles), le deuxième mot-clé le plus fréquent n'apparaissant que dans 9 documents (ISM :kinematics and dynamics). Peu des documents de cette classe sont donc décrits par les mêmes mots-clés, c'est pourquoi ils se retrouvent très dispersés dans la seconde classification. En effet, le seul mot-clé partagé par la quasi-totalité des documents de ce groupe (ISM : Bubbles) peut être rattaché à plusieurs grands thèmes, et les processus aléatoires de l'apprentissage peuvent favoriser des regroupements différents d'un apprentissage à l'autre ;
- les documents du cinquième groupe sont peu dispersés, et situés principalement dans une classe éloignées de la zone occupée par les documents des groupes précédents. Dans une classe adjacente, on retrouve les documents d'un sixième groupe, constituant la classe placée sous la cinquième classe dans la première classification (images absentes de la figure).

Nous observons environ 20% de classes de type peu homogène et instables que nous détectons par un comptage des documents dans la seconde classification. Pour cela, nous comptons les documents du groupe le plus dense auxquels nous ajoutons ceux des groupes adjacents. La classe est alors désignée comme instable si le nombre de ces documents est inférieur aux 2/3 (arbitraire) du nombre total des documents qu'elle comporte.

9.3 Conclusion

Nous avons décrit un exemple de classification d'articles obtenue avec une SOM, que nous avons explorée avec la carte bibliographique. Nous avons ainsi pu visualiser les zones relatives à certains thèmes et examiner leur contenu. Nous avons vérifié que cette classification a les caractéristiques que nous avons annoncées dans la partie 6.1 : un arrangement bidimensionnel des données tel que des données voisines aient des caractéristiques proches.

Nous avons ensuite vérifié que les classifications obtenues ne sont pas uniques. En effet, nous avons visualisé les principales différences que peuvent présenter les classifications issues de deux apprentissages similaires. Ces différences sont uniquement dues aux processus aléatoires qui interviennent durant l'apprentissage.

De manière générale, une classification comporte environ 20% de classes instables, qui n'apparaissent pas dans d'autres classifications. Ces classes font apparaître des ruptures de la conservation des relations de voisinage (corrélations spatiales) entre les classes d'une carte à une autre. Elles correspondent en général à des thèmes qui peuvent être rattachés à l'un ou l'autre des thèmes généraux. Les autres classes se retrouvent approximativement (avec de petites modifications) dans les autres classifications, tandis que les zones relatives aux thèmes généraux gardent globalement des positions relatives et des tailles comparables.