Méthode visuelle et interactive de partitionnement d'un ensemble de données à l'aide de graphes de voisinage construits par des fourmis artificielles

Julien Lavergne*, Hanane Azzag**
Christiane Guinot*,***, Gilles Venturini*

*Laboratoire d'Informatique,
Ecole Polytechnique de l'Université de Tours,
64 avenue Jean Portalis, 37200 Tours, France
{julien.lavergne,gilles.venturini}@univ-tours.fr,
http://www.antsearch.univ-tours.fr/webrtic
**Laboratoire d'Informatique de l'Université Paris-Nord XIII
CNRS - UMR 7030
99, avenue Jean-Baptiste Clément, 93430 Villetaneuse, France
hanane.azzag@lipn.univ-paris13.fr,
http://www-lipn.univ-paris13.fr/A3/
***CE.R.I.E.S, 20 rue Victor Noir, 92521 Neuilly-Sur-Seine, France
christiane.guinot@ceries-lab.com,
http://www.ceries.com

Résumé. Nous présentons dans cet article une méthode de découverte visuelle et interactive d'un partitionnement de données qui s'appuie sur la visualisation d'un graphe de voisinage obtenu à l'aide de la méthode biomimétique AntGraph. Le but est de construire ce graphe en un temps de calcul très faible grâce aux principes de l'heuristique développée, puis de laisser l'expert du domaine procéder de manière interactive à la définition d'une classification sur l'ensemble de données considéré. Nous proposons une évaluation expérimentale de notre méthode interactive sur un panel d'utilisateurs experts et non-experts et comparons les résultats obtenus en terme de qualité avec une méthode de classification interactive visuelle à base de points d'intérêts, et deux méthodes automatiques de classification que sont la Classification Ascendante Hiérarchique et AntTree. Nous montrons finalement que l'utilisation d'une technique de visualisation et d'exploration interactive par des utilisateurs experts ou non sur des graphes de données construits par AntGraph permet de découvrir une classification ayant une qualité similaire à celles obtenues avec des méthodes interactives ou automatiques.

1 Introduction

La classification est un des domaines importants de la fouille de données (Jain et al., 1999). La majorité des méthodes de classification sont automatiques. Cependant, dans de nombreux cas, les résultats obtenus nécessitent une validation de la part de l'expert du domaine. D'autre part, les méthodes automatiques fournissent peu d'informations sur les classes (e.g. densité, forme, proximité). Les informations généralement retournées concernent le nombre de classes et l'appartenance des données à ces classes. Des actions supplémentaires de la part de l'expert du domaine peuvent être alors nécessaires pour améliorer les résultats. A ce titre, l'utilisation partielle ou exclusive de méthodes interactives de classification (donc non automatiques) permet à l'expert du domaine d'agir sur la classification. Les méthodes de visualisation associées permettent, conjointement avec des techniques d'interaction, d'explorer les ensembles de données et d'extraire plus facilement de nouvelles caractéristiques des classes et données. Nous proposons dans cet article une adaptation de notre méthode de construction de graphes de voisinage AntGraph (Lavergne, 2008) à la fouille visuelle de données pour des tâches de classification interactive. A ce titre, nous montrons que notre méthode offre une qualité satisfaisante de partitionnement d'un ensemble de données et qu'elle permet la découverte des propriétés topologiques d'un ensemble de données.

La suite de cet article est organisée comme suit : dans la section 2, nous proposons un introduction aux tâches de classification en fouille visuelle de données. Dans la section 3, nous introduisons les approches existantes à base de graphes de voisinage et décrivons l'intérêt d'utiliser la méthode AntGraph pour produire une technique de représentation visuelle et d'exploration interactive d'un ensemble de données sous la forme d'un graphe de voisinage. Ceci dans le but de faciliter l'extraction de connaissances à partir des données et exploiter l'information issue des relations de voisinage. Nous détaillons ensuite, dans la section 4, la méthodologie employée pour évaluer la qualité de partitionnement réalisée par notre méthode de construction de graphes de voisinage en classification interactive (i.e. jugement de l'utilisateur). Nous proposons également, dans la section 5, plusieurs études comparatives avec d'autres méthodes de classification (interactive et automatiques). Finalement, nous concluons et énonçons les perspectives envisagées.

2 Introduction à la classification interactive

Il existe plusieurs méthodes de visualisation interactive dans un processus d'extraction de connaissances à partir des données, autrement appelé *Visual Data Mining* (VDM). Les nombreux avantages de la visualisation d'information associée à la fouille de données sont reconnues par la communauté (Fayyad et al., 2001), mais de nombreux points restent encore à résoudre (Chen, 2005). Les principales difficultés rencontrées concernent dans un premier temps la prise en compte de l'utilisateur au sein du processus de visualisation (Himberg, 2004), suivie de la nécessité de proposer de nouveaux modèles de visualisation adaptés (Card et al., 1999). Le dernier point concerne l'adéquation des méthodes de visualisation au niveau d'exigence de la discipline d'application concernée (e.g. marketing, sondage).

Dans le cas d'un problème de classification, les méthodes de VDM sont d'une grande utilité (Keim, 2002). Dans le cas supervisé, elles couplent les opérations interactives d'exploration et découverte de connaissances. Elles guident l'expert dans sa validation des résultats. Dans le cas non supervisé, elles permettent pleinement la construction de la classification et implicitement sa validation par l'utilisateur (desJardins et al., 2007). Les nombreuses méthodes disponibles utilisent pour la plupart des représentations symboliques et synthétiques de l'information (i.e. visages de (Chernoff, 1973), les arbres de décision (Ankerst et al., 1999), les scatterplots de (Becker et Cleveland, 1987), Multi Dimensional Scaling (MDS) (Faloutsos et Lin, 1995), les Support Vector Machines (SVM) (Caragea et al., 2005), les cartes auto-organisatrices (SOM) (Vesanto, 2002), Interactive Visual Clustering (desJardins et al., 2007)). Pour de plus amples informations, nous invitons à la lecture d'un état de l'art (Da Costa, 2007) portant sur la problématique de classification interactive en fouille visuelle de données.

Malheureusement, la plupart de ces méthodes traitent en majorité des données numériques et très peu les autres types (i.e. symbolique, textuel, mixte). En outre, un très faible nombre de ces méthodes considèrent les grands volumes de données (Keim et Ankerst, 2001; Fekete et Plaisant, 2002) (e.g. 1 million). Dans la majorité des cas, il est nécessaire de proposer des simplifications quant à la visualisation d'importantes quantités de données multidimensionnelles comme par exemple la sélection des dimensions et items les plus caractéristiques d'un ensemble (Sammon, 1969). De nombreuses méthodes et heuristiques existent en ce sens (Rossi, 2006). Cependant, peu de méthodes tiennent compte des relations de voisinage entre les données dans un processus de visualisation. En effet, les opérations de simplification (réduction) ne préservent pas toujours les relations de voisinage qui peuvent exister entre les données. Ainsi la visualisation qui en résulte est souvent faussée (ne rend pas compte de la réalité de l'information).

3 Classification visuelle interactive avec AntGraph

Nous nous sommes naturellement intéressés aux modèles de graphes de voisinage. A ce titre, nous rappelons l'existence d'un état de l'art sur ces méthodes dans (Hacid, 2008). Parmi ces modèles, nous pouvons citer les méthodes qui ont été utilisées pour l'indexation de documents (e.g. image (Ambauen et al., 2003; Hacid, 2008), texte (Clech et Zighed, 2004)). Les ensembles de données considérées peuvent atteindre quelques dizaines de milliers d'items pour les versions incrémentales de ces méthodes. Cependant, elles ne sont pas clairement adaptées au processus de visualisation d'un graphe avec le support des opérations de navigation et d'exploration interactive nécessaires à la fouille visuelle de données (Herman et al., 2000). Plus généralement, l'utilisation des méthodes classiques de construction de graphes de voisinage dans un processus de fouille visuelle de données s'appliquent majoritairement à l'exploration de bases de données

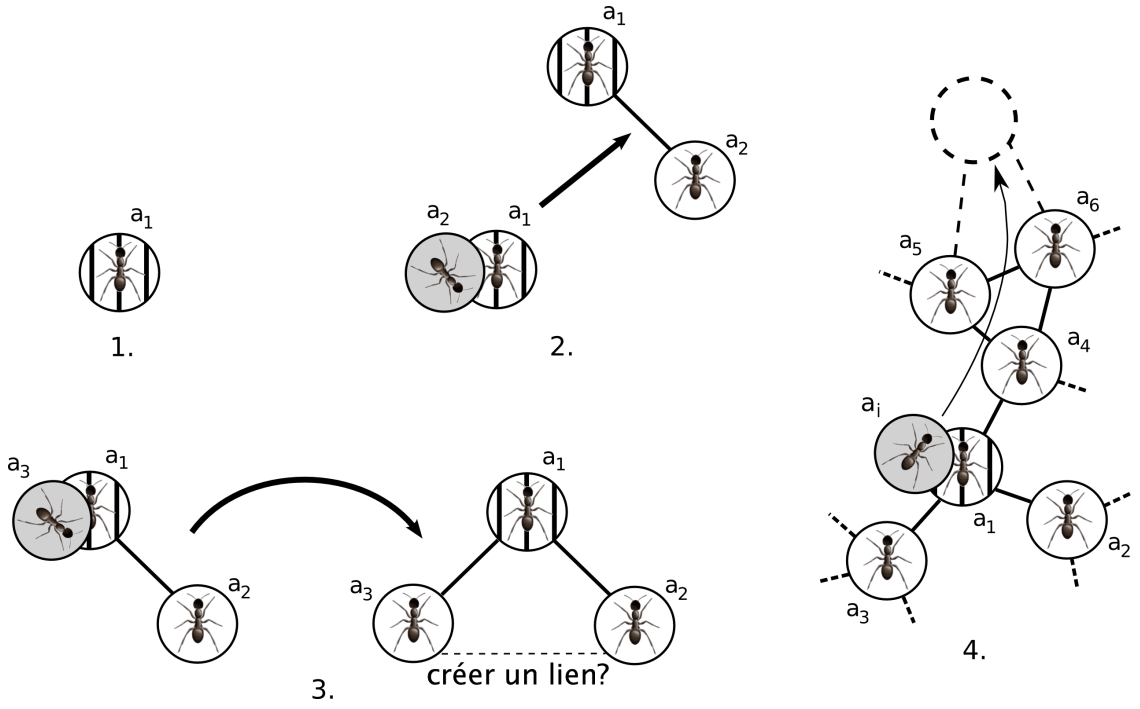
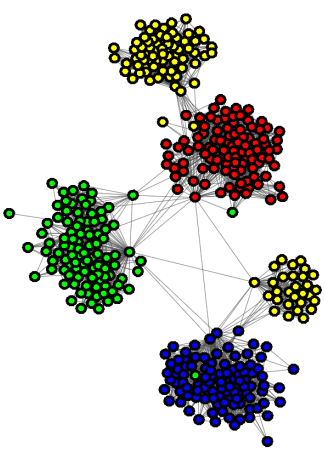


FIG. 1 – *Principes de construction d'un graphe avec les fourmis artificielles. Les étapes 1, 2 et 3 illustrent le début de la construction d'un graphe de proximité tandis que l'étape 4 généralise cette construction avec une nouvelle fourmi a_i introduite dans le graphe. a_i se déplace de fourmi en fourmi (i.e. la fourmi parcourt le chemin de similarité locale maximum).*

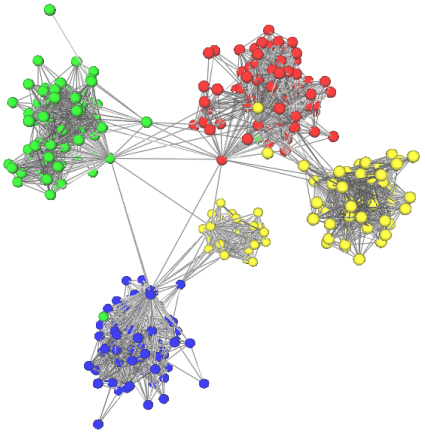
spatiales (Lee, 2002) ou de bases relationnelles pré-existantes (rendre compte visuellement des relations existantes entre entités d'une base). Nous pouvons citer les exemples suivants de graphes de voisinage orientés Web (ayant une interactivité avec l'utilisateur) : TouchGraph (TouchGraph LLC, 2006) pour la visualisation générique et dynamique d'un graphe de documents pré-construits, Audiomap (TuneGlue, 2006) pour la découverte de la proximité d'un artiste de musique avec d'autres, ThinkBase (Hirsch et al., 2008) une nouvelle manière d'interroger et explorer le Web ou encore VisuWords (Octopus, 2007) qui permet d'explorer la proximité sémantique d'un terme avec d'autres termes et concepts.

Nous proposons dans cet article un nouveau modèle de représentation visuelle interactive qui est une adaptation de notre méthode de construction incrémentale de graphes de voisinage AntGraph à de la classification visuelle interactive non supervisée. Notre méthode construit un graphe de proximité avec des fourmis artificielles à partir d'un ensemble de données (Lavergne et al., 2007). A l'initialisation, nous considérons un ensemble de données (fourmis) sous la forme d'un flux (dont l'ordre des données n'est pas connu a priori) et nous choisissons la première fourmi du flux (notée a_1) comme support fixe et noeud d'entrée du graphe. Ensuite, tant qu'il y a des fourmis non connectées, nous considérons une fourmi a_i qui est insérée dans le graphe en a_1. Cette fourmi se déplace de fourmi en fourmi jusqu'à se connecter sur celle qui lui est la plus similaire localement. Elle parcourt pour cela le chemin de plus grande similarité (voir figure 1). Les fourmis optimisent localement leur déplacement et peuvent ainsi traverser rapidement de grands ensembles de noeuds.

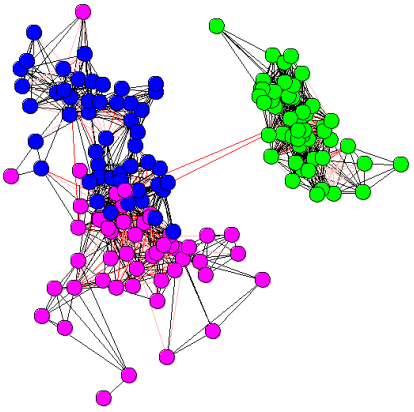
Nous visualisons ensuite nos graphes à l'aide d'une méthode à base de forces (Fruchterman et Reingold, 1991). L'approche utilisée définit un ensemble de forces qui sont exercées entre les noeuds du graphe : des forces de répulsion entre tous les noeuds tandis que des forces d'attraction sont appliquées sur les noeuds/données adjacents/similaires. La méthode itère, se stabilise et nous bénéficions finalement d'une disposition esthétique des noeuds du graphe. Cette méthode nous permet également de préciser une distance géométrique entre les noeuds relative de la distance réelle entre les données. Nous remarquons ainsi visuellement la proximité des noeuds/données similaires. Nous visualisons également la structure du graphe avec une distinction de la forme globale des classes et disposons d'une navigation guidée par le contenu avec perception de la proximité réelle entre les données. Les figures 2(a), 2(b), 2(c) et 2(d) sont des exemples



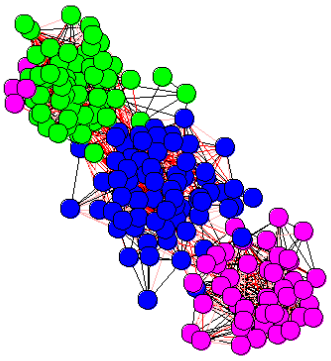
(a) visualisation 2D d'un graphe construit sur la base artificielle ART6.



(b) visualisation 3D du graphe 2(a).



(c) visualisation 2D d'un graphe construit sur la base IRIS.



(d) visualisation 2D d'un graphe construit sur la base WINE.

FIG. 2 – *Exemples de visualisations de graphes construits avec notre méthode AntGraph sur quelques bases artificielles et réelles (Monmarché, 2000; Blake et Merz, 1998)* .

de graphes construits avec notre méthode à partir de bases classiques et artificielles. L'aspect incrémental de notre méthode est fondamental et nous permet de considérer un ensemble de données sous la forme d'un flux. Il nous est possible de compléter la construction/visualisation d'un graphe au cours du temps sans contraindre et gêner l'expert du domaine dans son exploration interactive de la structure. Les nouveaux noeuds/données apparaissent dans le graphe et prennent place dans la visualisation.

Dans le cadre de la classification interactive en fouille visuelle de données, nous avons développé un outil de visualisation et d'exploration interactive de graphes de voisinage. Il implémente l'algorithme AntGraph pour la construction d'un graphe de voisinage à partir d'un ensemble de données. Le graphe ainsi construit est ensuite visualisé avec la méthode à base de forces décrite précédemment. L'utilisateur profite ainsi d'une visualisation rendant compte d'un partitionnement des données en clusters distincts. Nous sommes également motivés par le fait que nous souhaitons utiliser au mieux les interactions fournies par notre outil de classification visuelle interactive et ainsi augmenter les possibilités d'extraction et d'interprétation des connaissances contenues dans un graphe par l'utilisateur expert du domaine, voire non-expert. Au sein d'une même visualisation, les informations sont diverses et nombreuses. Avec notre outil, l'utilisateur est ainsi capable d'obtenir :

- le nombre de classes. D'un point de vue interactif, le nombre de classes dépend des choix effectués par l'utilisateur (sélection et étiquetage de groupes de noeuds/données),

- les relations entre classes (i.e. proximité, imbrications),
- la forme des classes :
  - découvrir de manière visuelle dans son intégralité la partition de l'ensemble de données en différents clusters ainsi que localement la périphérie et le coeur des groupes de données. Il doit être ainsi possible de passer du point de vue global à une sous-partie (i.e. observer un/plusieurs clusters). A titre d'exemple au niveau global, nous pouvons citer la découverte de la densité d'un cluster ou bien la disparité des données au sein d'un groupe,
  - identifier au niveau local les données isolées ou fortement connectées et pouvoir considérer une ou plusieurs données en particulier et leurs relations,
- sélectionner de manière interactive les nœuds/données ou groupes de données.

Avec cette dernière interaction, nous pouvons enregistrer la classification réalisée par l'utilisateur, ce qui va nous permettre ensuite de réaliser les différentes mesures pour évaluer sa qualité.

Finalement, notre outil dispose des opérations interactives nécessaires (e.g. sélection de noeuds, divers zooms, différents mouvements de rotations et plusieurs types de caméras) pour réaliser une classification interactive des données d'un graphe. Il propose bien une visualisation avec navigation et exploration interactive des graphes construits avec notre méthode. L'information relationnelle contenue au sein d'un graphe (liens avec rendu de la proximité réelle entre les noeuds/données) permet de guider l'utilisateur dans ses choix. Nous proposons deux modes de visualisation (2D et 3D) à l'utilisateur pour effectuer une classification interactive. L'utilisateur peut être, dans le cas de notre outil, expert du domaine ou non. Nous insistons principalement ci-dessous sur la 3D qui permet de lever des ambiguïtés (i.e. occlusion de noeuds/données) au niveau de la visualisation (Tory et al., 2006). La figure 3 illustre, à titre d'exemple, le cas complet d'une classification interactive en 3D sur un graphe de voisinage construit avec notre méthode AntGraph. Nous proposons également la variante 2D illustrée par la figure 4.

4 Expérimentations

4.1 Protocole et jeux de données

Nous souhaitons montrer l'efficacité de notre méthode en terme de classification non supervisée des données. Comme expliqué dans la section précédente, nous construisons un graphe avec la méthode AntGraph à partir d'un ensemble de données. Une fois le graphe construit, nous évaluons la qualité de partitionnement en terme de classification visuelle interactive et comparons les résultats obtenus avec ceux de 3 autres méthodes de classification : 1) une méthode de classification visuelle interactive à base de points d'intérêts (i.e. POID3D1 (Da Costa, 2007)) ; 2) une méthode de classification automatique classique et reconnue, la CAH (Lance et Williams, 1967; Sneath et Sokal, 1973) et une méthode biomimétique, AntTree (Azzag, 2005).

Pour l'ensemble des tests que nous avons réalisés avec notre méthode, nous avons décidé d'évaluer la classification interactive obtenue à la fois en terme de nombre de classes trouvées (C_T), de pureté de classification (P_R) ainsi que d'erreur couple (E_C) (également appelé indice de Rand). Nous réalisons les tests sur des bases dont nous connaissons les caractéristiques (e.g. nombre de classes réelles) mais que nous n'utilisons pas dans les tests.

Nous avons réalisé plusieurs études comparatives sur des ensembles de données numériques : 6 jeux artificiels (Monmarché, 2000) et 8 bases classiques du UCI Repository of Machine Learning (Blake et Merz, 1998) et du CE.R.I.E.S. (Guinot et al., 2004). Nous considérons ainsi les bases artificielles présentées qui ont été générées avec une loi uniforme (i.e. différents nombres de classes, avec et sans recouvrement). Les bases réelles présentent également des difficultés bien identifiées. Cependant, le cadre de nos évaluations concerne la classification interactive non supervisée et nous n'utiliserons ces connaissances que pour éventuellement pondérer nos remarques sur les résultats obtenus.

Pour évaluer au mieux la qualité du partitionnement proposé par notre méthode sous la forme d'un graphe, nous considérons un tri aléatoire des données pour chaque test de classification interactive réalisé par un utilisateur (expert ou non-expert). Nous montrons, comme nous l'avons fait dans (Lavergne, 2008), qu'un tri aléatoire des données n'influence que très peu la qualité des graphes construits et n'empêche pas une bonne classification interactive (voir sections 5.1 et 5.2).

Nous comparons, ensuite, les méthodes sur un même ensemble de jeux de données en terme de classification interactive. Cela nécessite l'intervention de l'utilisateur. Ce dernier peut être expert du domaine.

Pour évaluer au mieux notre méthode, nous avons réalisé des tests avec les deux types d'utilisateurs. Nous présentons ci-dessous les deux profils (expert et non-expert).

4.2 Point de vue : expert du domaine

L'expert du domaine, en l'occurrence, est l'un des auteurs de cet article. Il connaît bien le fonctionnement de la méthode de construction AntGraph ainsi que les caractéristiques des ensembles de tests considérés sur lesquels la méthode est évaluée ci-après. Cependant, dans un cadre non supervisé, comme c'est ici le cas, l'expert du domaine ne dispose d'aucune information (e.g. la classe) permettant d'identifier les données. L'information des classes réelles n'est bien sûr pas du tout utilisée lors des tests.

Afin de ne pas présenter des résultats qui pourraient être qualifiés de biaisés (connaissances de l'expert du domaine), nous avons également choisi de réaliser une évaluation par des utilisateurs non-experts. Nous rendons également compte d'une accessibilité de notre méthode dans le cadre de la fouille visuelle de données.

4.3 Point de vue : utilisateurs non-experts

Nous avons constitué un panel de 8 utilisateurs que nous avons voulu de catégories socio-professionelles et d'âges différents (i.e. 4 personnes entre 20-30 ans, étudiant(e)s en psychologie, informatique et lettres ; 2 personnes entre 30-50 ans, personnes actives dans le privé, domaines du bâtiment et de l'immobilier ; 2 personnes retraitées, entre 60 et 65 ans).

Toutes ces personnes ont été sélectionnées avec comme aptitude minimale l'utilisation courante d'un ordinateur. Nous nous sommes assurés que ces personnes ne connaissaient pas les ensembles de données ni leurs caractéristiques. Nous cherchons à montrer que notre méthode fonctionne également avec des utilisateurs non-experts et offre des résultats comparables voire identiques.

Nous avons présenté l'outil que nous avons développé aux utilisateurs non-experts et leur avons expliqué le maniement de l'outil. Nous avons réalisé une phase d'apprentissage et nous nous sommes assurés que chaque utilisateur maîtrise pleinement notre outil de visualisation et d'exploration interactive d'un graphe de données.

Un utilisateur non-expert se doit d'être au minimum capable de :

- visualiser le graphe de données,
- manipuler l'outil de navigation et d'exploration interactive,
- maîtriser les opérations interactives proposées (zoom avec distorsion, déplacement avec vue à la première personne, caméra à coordonnées sphériques, outil de sélection des données),
- identifier les groupes de données,
- sélectionner les données par groupe.

En résumé, l'utilisateur a pour but de visualiser chaque graphe de données construit par la méthode AntGraph, d'identifier visuellement les groupes de données (clusters) en son sein, de les sélectionner avec l'opération interactive adéquate. Pour chaque test, une fois le graphe entièrement étiqueté par l'utilisateur, notre outil évalue le taux de bonne classification (P_R) et le nombre de classes trouvées (C_T).

5 Etudes comparatives

Nous présentons, dans cette section, les études comparatives en classification interactive, énoncées dans la section précédente, de notre approche avec une méthode visuelle interactive (Da Costa, 2007) et deux méthodes de classification automatique. La première méthode automatique, la Classification Ascendante Hiérarchique (CAH, (Lance et Williams, 1967; Sneath et Sokal, 1973)), est une méthode classique tandis que la seconde méthode est qualifiée de biomimétique. Il s'agit de la méthode AntTree de (Azzag, 2005) qui réalise une classification non supervisée d'un ensemble de données sous la forme d'une hiérarchie de fourmis/données.

5.1 Comparaison avec une méthode visuelle de classification interactive

Nous avons décidé de valider de manière interactive les capacités de classification de notre méthode AntGraph par rapport à une autre méthode de classification visuelle interactive : celle des POints d'Intérêts (POI), développée par (Da Costa, 2007), pour laquelle nous considérons POI3D1. AntGraph et POI3D1 réalisent bien une classification interactive non supervisée, utilisent une mesure de similarité, ce qui permet finalement de considérer les mêmes ensembles de données pour l'étude comparative.

Nous présentons dans cette section les résultats obtenus par notre expert et un panel d'utilisateurs non-experts avec notre méthode en comparaison avec ceux obtenus sur POI3D1, sur un ensemble de 6 bases réelles numériques (Blake et Merz, 1998) (mêmes ensembles de bases que dans (Da Costa, 2007)). Le tableau 1 regroupe les résultats obtenus par notre expert sur la méthode AntGraph ainsi que ceux obtenus par l'expert du domaine avec la méthode POI3D1. Nous présentons également, dans ce même tableau, un extrait des résultats que nous avons obtenus en visualisation 2D par notre expert. Le tableau 2 contient, quant à lui, les résultats de notre panel d'utilisateurs non-experts obtenus sur le même ensemble de 6 bases de données aussi bien en 2D qu'en 3D. Nous rappelons que nous disposons d'un panel de 8 utilisateurs (voir section 4.3). 96 sessions d'étiquetage, par des utilisateurs non-experts, des noeuds/données de graphes construits avec notre méthode ont été ainsi réalisées. Les résultats correspondants de la méthode POI3D1 (voir tableau 2) sont issus d'un panel d'utilisateurs non-experts sélectionnés dans le cadre des travaux de thèse de (Da Costa, 2007).

Notre méthode propose pour la majorité des bases testées une bonne classification (i.e. 4 bases sur 6 avec une $P_R \geq 89\%$ pour l'expert, $P_R \geq 88\%$ pour les utilisateurs non-experts). Nous pouvons ajouter que les résultats obtenus en terme de pureté par nos utilisateurs non-experts sont proches de ceux obtenus par l'expert du domaine sauf pour les bases bruitées (i.e. PIMA et VEHICLE) où les résultats de l'expert sont de meilleure qualité. Nous remarquons également que les taux de bonne classification (P_R) pour notre méthode sont très similaires à ceux obtenus avec la méthode par points d'intérêts POI3D1 chez les utilisateurs non-experts. Ainsi sur les bases faiblement bruitées ou pas du tout, notre méthode donne des résultats en terme de pureté équivalents à ceux de POI3D1 voire meilleurs. Quant aux résultats obtenus sur AntGraph en terme de classes trouvées (C_T) aussi bien par l'expert que les non-experts, ces derniers sont sensiblement plus élevés que ceux de POI3D1 mais restent proches du nombre de classes réelles C_R. Rappelons que les résultats obtenus dans le cadre de notre méthode sont ceux d'une méthode interactive où l'utilisateur peut découvrir autant de clusters qu'il le désire selon son jugement. Il peut avoir tendance à sélectionner de plus petits groupes (clusters) de données, estimant des sous-groupes de données similaires entre elles. Finalement, avec les résultats présentés, nous pouvons conclure que notre méthode de classification visuelle interactive de données sur des graphes de voisinage réalise dans l'ensemble une classification satisfaisante. Elle est cependant plus sensible au bruit contenu au sein des données que POI3D1.

Nous comparons maintenant notre méthode à deux méthodes automatiques de classification non supervisée.

5.2 Comparaison avec deux méthodes de classification automatique

Nous avons testé chacune des méthodes sur un ensemble de 14 jeux de données : 6 bases artificielles (Monmarché, 2000) et 8 bases classiques (Blake et Merz, 1998; Guinot et al., 2004). Nous mesurons, pour chaque test de méthode sur un jeu de données, la pureté de classification P_R, l'Erreur Couple E_C, ainsi que le nombre de classes trouvées C_T (nombre de sélections de groupes de données réalisées par l'utilisateur). Dans le cas de notre méthode, nous considérons les résultats obtenus en classification visuelle interactive 3D par l'expert du domaine sur les 14 bases. Il s'agit des résultats pour l'ensemble des 6 bases considérées dans la section 5.1 complétés de résultats supplémentaires pour les 8 bases restantes (les 6 bases artificielles ART{1,..,6} et les bases CERIES et GLASS).

La Classification Ascendante Hiérarchique (CAH, (Lance et Williams, 1967; Sneath et Sokal, 1973)) est la première méthode de classification automatique à laquelle nous comparons notre méthode. Nous avons choisi pour fixer le nombre de classes trouvées (C_T), lors des tests sur la CAH, la valeur du saut maximal du critère de Ward (Azzag, 2005). Le tableau 3 contient l'ensemble des résultats obtenus pour notre méthode et la CAH.

Bases	N	M	C_R	AntGraph évalué en 2D				AntGraph évalué en 3D				POI3D1	
				C_T	σ_{C_T}	P_R	σ_{P_R}	C_T	σ_{C_T}	P_R	σ_{P_R}	C_T	P_R
IRIS	150	4	3	3,7	[0,8]	0,88	[0,04]	3,9	[0,78]	0,92	[0,03]	3	0,87
PIMA	768	8	2	10,2	[0,6]	0,70	[0,03]	10,8	[0,74]	0,71	[0,03]	2	0,65
SOYBEAN	47	35	4	4,5	[0,7]	0,92	[0,03]	4,8	[0,86]	0,96	[0,02]	3	0,75
THYROID	215	5	3	3,9	[0,8]	0,85	[0,04]	4,1	[0,85]	0,89	[0,03]	3	0,84
VEHICLE	846	18	4	6,9	[0,7]	0,44	[0,02]	7,3	[0,83]	0,46	[0,02]	4	0,75
WINE	178	12	3	4,1	[0,7]	0,89	[0,03]	4,2	[0,60]	0,91	[0,02]	3	0,86

TAB. 1 – *Résultats obtenus par un utilisateur expert sur la méthode AntGraph et un autre utilisateur expert sur POI3D1 (Da Costa, 2007) sur des bases numériques (Blake et Merz, 1998). N représente le nombre de données, M le nombre d'attributs et C_R le nombre de classes réelles d'une base.*

Bases	N	M	C_R	AntGraph évalué en 2D				AntGraph évalué en 3D				POI3D1	
				C_T	σ_{C_T}	P_R	σ_{P_R}	C_T	σ_{C_T}	P_R	σ_{P_R}	C_T	P_R
IRIS	150	4	3	3,5	[0,7]	0,86	[0,03]	3,6	[0,80]	0,90	[0,02]	3	0,90
PIMA	768	8	2	9,8	[0,7]	0,53	[0,04]	9,6	[0,78]	0,50	[0.03]	2	0,68
SOYBEAN	47	35	4	4,3	[0,8]	0,91	[0,03]	5,0	[0,89]	0,93	[0,05]	4	0,98
THYROID	215	5	3	4,1	[0,9]	0,83	[0,05]	4,3	[0,89]	0,88	[0,04]	3	0,90
VEHICLE	846	18	4	7,5	[0,8]	0,33	[0,03]	8,3	[1,07]	0,31	[0,05]	3	0,78
WINE	178	12	3	4,2	[0,7]	0,86	[0,03]	4,0	[0,63]	0,88	[0,02]	3	0,81

TAB. 2 – *Résultats obtenus par un panel de 8 utilisateurs non-experts sur la méthode AntGraph et par un autre panel d'utilisateurs non-experts sur POI3D1 (Da Costa, 2007) sur des bases numériques (Blake et Merz, 1998).*

Dans l'ensemble, notre méthode propose une répartition homogène des données dans les classes trouvées (faibles valeurs d'E_C) avec une qualité équivalente à la CAH. Finalement, notre méthode propose une qualité de classification satisfaisante sur l'ensemble des bases testées mis à part le cas des bases bruitées là où même une méthode de classification automatique telle que la CAH ne fait pas mieux (i.e. bases VEHICLE et PIMA).

Nous comparons maintenant les résultats de notre méthode à ceux obtenus avec une méthode biomimétique de classification automatique (AntTree, (Azzag, 2005)) sur ce même ensemble de 14 bases (i.e. artificielles et classiques). Du fait de résultats volumineux et par soucis de place, nous proposons une synthèse des résultats obtenus dans le tableau 3. Notre méthode obtient des résultats très similaires à ceux obtenus par la méthode AntTree. Sur l'ensemble des bases testées, les résultats en terme de classification interactive sur des graphes de données sont légèrement meilleurs. Les deux méthodes héritant du même principe de construction (i.e. arbre dans un cas, graphe dans l'autre), nous pouvons statuer qu'en terme de classification interactive, l'interaction de l'utilisateur sur un graphe construit avec notre méthode donne des résultats satisfaisants proche d'une méthode biomimétique de classification automatique telle que AntTree.

6 Conclusions et perspectives

Nous avons présenté dans cet article une méthode de découverte visuelle et interactive d'un partitionnement de données obtenue à partir de la visualisation d'un graphe de voisinage construit avec la méthode AntGraph. Nous avons ainsi pu mettre en évidence les capacités de classification interactive de notre méthode sur la base d'un graphe de voisinage construit à partir d'un ensemble de données en un temps très court. Nous démontrons les propriétés de notre méthode en nous appuyant sur trois études comparatives (i.e. POI3D1, CAH et AntTree). Les résultats que nous avons obtenus mettent en évidence une bonne qualité de classification de notre méthode sur l'ensemble des bases testées, voire identique avec les méthodes automatiques.

L'intérêt de notre méthode de construire des graphes où les données similaires sont regroupées en clusters et distantes selon leur proximité réelle, offre à l'utilisateur dans un contexte de fouille visuelle de données une aide précieuse en terme d'extraction et interprétation des connaissances à partir des données.

Bases	C_R	AntGraph évalué en 3D par l'expert						CAH			AntTree		
		E_C	σ_{E_C}	C_T	σ_{C_T}	P_R	σ_{P_R}	E_C	C_T	P_R	E_C	C_T	P_R
ART1	4	0,20	[0,02]	8,2	[0,71]	0,82	[0,06]	0,15	5	0,84	0,19	5,20	0,77
ART2	2	0,21	[0,05]	4,8	[0,86]	0,96	[0,02]	0,14	3	0,98	0,24	4,93	0,95
ART3	4	0,21	[0,01]	5,9	[0,82]	0,90	[0,01]	0,16	3	0,89	0,25	5,53	0,86
ART4	2	0,27	[0,05]	5,1	[0,63]	0,96	[0,03]	0,13	3	1,00	0,28	5,53	0,99
ART5	9	0,11	[0,01]	9,2	[0,96]	0,62	[0,03]	0,08	10	0,78	0,17	6,00	0,50
ART6	4	0,09	[0,04]	5,0	[0,60]	0,92	[0,02]	0,03	5	1,00	0,07	6,40	0,97
IRIS	3	0,09	[0,02]	3,9	[0,78]	0,92	[0,03]	0,13	3	0,88	0,18	3,93	0,84
PIMA	2	0,48	[0,02]	10,8	[0,74]	0,71	[0,03]	0,48	3	0,65	0,48	13,47	0,72
SOYBEAN	4	0,06	[0,00]	4,8	[0,86]	0,96	[0,02]	0,08	6	1,00	0,10	6,67	0,99
THYROID	3	0,29	[0,04]	4,1	[0,85]	0,89	[0,03]	0,35	5	0,84	0,27	5,93	0,87
VEHICLE	4	0,32	[0,01]	7,3	[0,83]	0,46	[0,02]	0,39	3	0,35	0,31	8,47	0,45
WINE	3	0,16	[0,02]	4,2	[0,60]	0,91	[0,02]	0,20	6	0,84	0,21	8,20	0,88
CERIES	6	0,16	[0,02]	9,4	[2,24]	0,71	[0,07]	0,24	3	0,56	0,15	9,53	0,75
GLASS	6	0,32	[0,02]	8,1	[0,67]	0,55	[0,03]	0,43	3	0,49	0,34	8,07	0,55

TAB. 3 – *Comparaison des résultats obtenus, en classification interactive, entre la méthode AntGraph, la CAH (Lance et Williams, 1967; Sneath et Sokal, 1973) et la méthode AntTree (Azzag, 2005). Ensemble de bases de données de (Monmarché, 2000; Blake et Merz, 1998; Guinot et al., 2004). C_R, C_T, P_R, M, N et E_C représentent respectivement le nombre de classes réelles, le nombre de classes trouvées, la pureté de classification, le nombre d'attributs, le nombre de données et l'indice de Rand pour une base.*

Parmi les perspectives que nous envisageons pour la suite de nos travaux, nous pourrions étendre notre méthode de visualisation à l'aide d'un couplage avec une méthode à base de points d'intérêts pour optimiser le placement des noeuds dans l'espace de représentation visuelle. Une autre perspective serait de pouvoir visualiser dans son ensemble de manière interactive le graphe construit sur un grand volume de données (e.g. 1 million de données) avec l'extension de notre méthode, qui consiste à construire un grand graphe hiérarchique de voisinage (Lavergne, 2008). Nous souhaitons également compléter notre étude en classification interactive 2D auprès des utilisateurs non-experts. A ce titre, nous pourrions envisager d'enrichir les études précédentes à l'aide de résultats comparatifs avec d'autres méthodes interactives. Nous pourrions également nous confronter à de nouveaux protocoles en Interface Homme Machine et qualifier plus finement les apports de notre méthode en terme de fouille visuelle de données pour une problématique de classification non supervisée.

A terme, nous souhaitons proposer une méthode de visualisation interactive 2D et 3D de la topologie d'un ensemble de données (i.e. graphe de voisinage) performante pour les grands volumes de données (i.e. 1 million) et bénéficiant d'une expérience utilisateur enrichie.

Références

Ambauen, R., S. Fischer, et H. Bunke (2003). Graph edit distance with node splitting and merging and its application to diatom identification. In *Proceedings of the 4th International Workshop on Graph Based Representations in Pattern Recognition*, Volume 2726 of *Lecture Notes in Computer Science*, pp. 95–106.

Ankerst, M., C. Elsen, M. Ester, et H.-P. Kriegel (1999). Visual classification : an interactive approach to decision tree construction. In *KDD '99 : Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, pp. 392–396. ACM.

Azzag, H. (2005). *Classification hiérarchique par des fourmis artificielles : applications à la fouille de données et de textes pour le Web*. Ph. D. thesis, Université de Tours.

Becker, R. A. et W. S. Cleveland (1987). Brushing scatterplots. *Technometrics 29*(2), 127–142.

Blake, C. et C. Merz (1998). UCI repository of machine learning databases.

Caragea, D., D. Cook, et V. Honavar (2005). Visual methods for examining support vector machine results with applications to gene expression data analysis. Technical report, ISU Technical Report, Ames, I.

Card, S. K., J. D. Mackinlay, et B. Shneiderman (1999). *Readings in information visualization : using vision to think*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Chen, C. (2005). Top 10 unsolved information visualization problems. *IEEE Comput. Graph. Appl. 25*(4), 12–16.

Chernoff, H. (1973). Using faces to represent points in k–dimensional spae graphically. *Journal of the American Statistical Association 68*, 361–368.

Clech, J. et D.-A. Zighed (2004). *Document numérique*, Volume 8, Chapter Une technique de réétiquetage dans un contexte de catégorisation de textes, pp. 55–69. Lavoisier, Cachan.

Da Costa, D. (2007). *Visualisation et fouille interactive de données à base de points d'intérêts*. Ph. D. thesis, Université François Rabelais de Tours.

desJardins, M., J. MacGlashan, et J. Ferraioli (2007). Interactive visual clustering. In *IUI '07 : Proceedings of the 12th international conference on Intelligent user interfaces*, New York, NY, USA, pp. 361–364. ACM.

Faloutsos, C. et K.-I. Lin (1995). Fastmap : a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *SIGMOD Rec. 24*(2), 163–174.

Fayyad, U., G. G. Grinstein, et A. Wierse (2001). *Information Visualization in Data Mining and Knowledge Discovery*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Fekete, J.-D. et C. Plaisant (2002). Interactive information visualization of a million items. In *INFOVIS '02 : Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02) : Proceedings of the IEEE Symposium on Information Visualization (InfoVis'02)*, Washington, DC, USA, pp. 117. IEEE Computer Society.

Fruchterman, T. et E. Reingold (1991). Graph drawing by force-directed placement. *Software – Practice Experience 21*(11), 1129-1164.

Guinot, C., D. J.-M. Malvy, F. Morizot, M. Tenenhaus, J. Latreille, S. Lopez, E. Tschachler, et L. Dubertret (2004). *Classification of healthy human facial skin*. CRC Press. Textbook of Cosmetic Dermatology Third edition.

Hacid, H. (2008). *Un Environnement Informatique pour l'Interrogation et l'Accès Intelligent aux Bases de Données Complexes*. Ph. D. thesis, Université Lumière Lyon II.

Herman, I., G. Melançon, et M. S. Marshall (2000). Graph visualization and navigation in information visualization : A survey. *IEEE Transactions on Visualization and Computer Graphics 6*(1), 24–43.

Himberg, J. (2004). *From Insights to Innovations : Data Mining, Visualization, and User Interfaces*. Ph. D. thesis, Helsinki University of Technology, Espoo, Finland.

Hirsch, C., J. C. Grundy, et J. Hosking (2008). Thinkbase : A visual semantic wiki. In C. Bizer et A. Joshi (Eds.), *International Semantic Web Conference (Posters & Demos)*, Volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jain, A. K., M. N. Murty, et P. J. Flynn (1999). Data clustering : a review. *ACM Computing Surveys 31*(3), 264–323.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics 8*(1), 1–8.

Keim, D. A. et M. Ankerst (2001). Visual data mining and exploration of large databases. In *the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Freiburg, Germany.

Lance, G. et W. Williams (1967). A general theory of classificatory sorting strategies : I. hierarchical systems. *Computer journal 9*(4), 373–380.

Lavergne, J. (2008). *Algorithmes de fourmis artificielles pour la construction incrémentale et la visualisation interactive de grands graphes de voisinage*. Ph. D. thesis, Université François-Rabelais de Tours.

Lavergne, J., H. Azzag, C. Guinot, et G. Venturini (2007). Incremental construction of neighborhood graphs using the ants self-assembly behavior. *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007) 1*, 399–406.

Lee, I. (2002). *Multi-Purpose Boundary-Based Clustering on Proximity Graphs for Geographical Data Mining*. Ph. D. thesis, The University of Newcastle, Callaghan NSW 2308, Australia.

Monmarché, N. (2000). *Algorithmes de fourmis artificielles : applications à la classification et à l'optimisation*. Ph. D. thesis, Université de Tours.

Octopus, T. L. (2007). Visuwords : Online graphical dictionary http ://www.visuwords.com.

Rossi, F. (2006). Visual data mining and machine learning. In *Proceedings of XIVth European Symposium on Artificial Neural Networks (ESANN 2006)*, Bruges (Belgium), pp. 251–264.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput. 18*(5), 401–409.

Sneath, P. H. et R. R. Sokal (1973). *Numerical Taxonomy*. San Francisco : W.H. Freeman.

Tory, M., A. E. Kirkpatrick, et M. S. Atkins (2006). Visualization task performance with 2d, 3d, and combination displays. *IEEE Transactions on Visualization and Computer Graphics 12*(1), 2–13. Member-Torsten Moller.

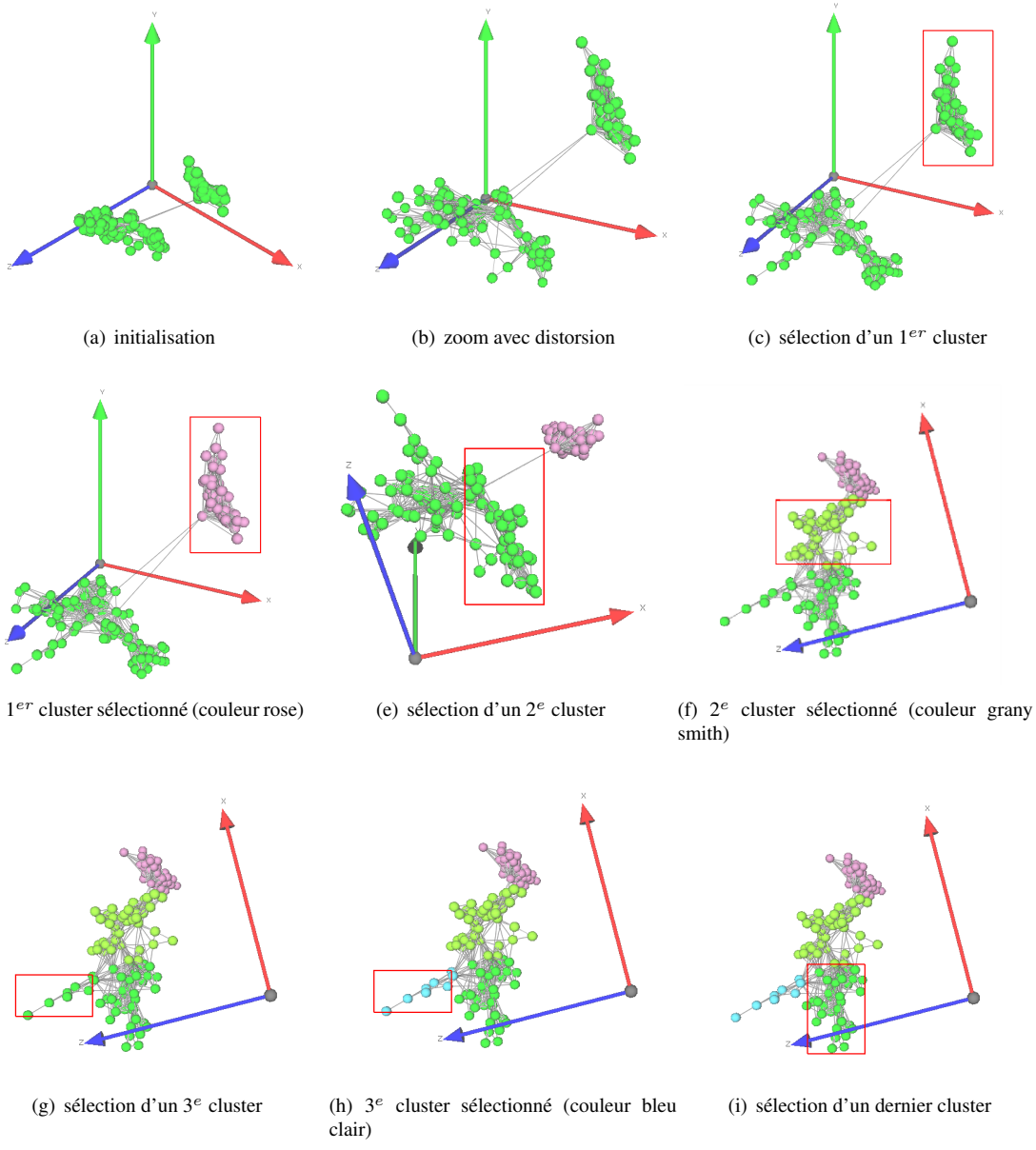
TouchGraph LLC (2006). Touchgraph navigator, http :/www.touchgraph.com.

TuneGlue (2006). Audiomap, http ://audiomap.tuneglue.net/.

Vesanto, J. (2002). *Data Exploration Process Based on the Self–Organizing Map*. Ph. D. thesis, Helsinki University of Technology, Espoo, Finland.

Summary

We introduce in this paper a method of visual discovery and interactive partitioning of a data set, which relies on visualizing a neighborhood graph obtained using the biomimetic method of proximity graphs construction, AntGraph (Lavergne, 2008). The goal is to build this graph with a short execution time thanks to principles of our building heuristic, and then let the domain expert interactively proceed to the definition of a clustering of the data set. We present an evaluation of our method with a panel of different users (i.e. domain expert and several non-experts), and we compare our tool in terms of quality to a visual interactive clustering method based on points of interest (Da Costa, 2007), and to two automatic clustering methods namely the HAC (Lance et Williams, 1967; Sneath et Sokal, 1973) and the biomimetic method, AntTree (Azzag, 2005). We finally show that the use of a visual technique, for interactive exploration by users of proximity graphs of data built by AntGraph, can confirm a quality of clustering with similar performance to those obtained with automatic and interactive methods.

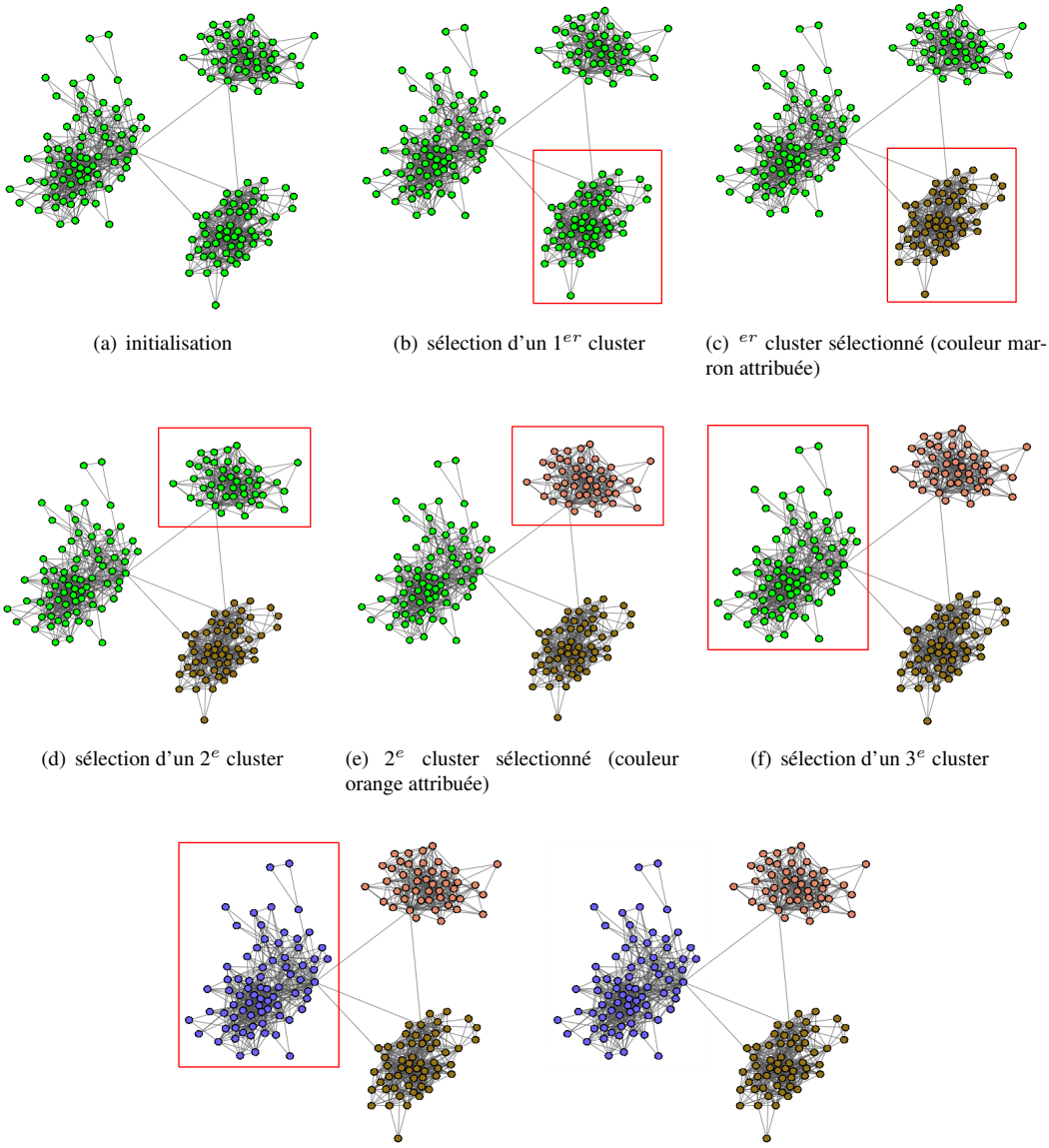


(a) initialisation (b) zoom avec distorsion (c) sélection d'un 1^{er} cluster

(d) 1^{er} cluster sélectionné (couleur rose) (e) sélection d'un 2^{e} cluster (f) 2^{e} cluster sélectionné (couleur grany smith)

(g) sélection d'un 3^{e} cluster (h) 3^{e} cluster sélectionné (couleur bleu clair) (i) sélection d'un dernier cluster

FIG. 3 – *Exemple d'une classification visuelle et interactive en 3D sur un graphe construit avec notre méthode AntGraph à partir de la base IRIS (150 données, 3 classes dont 2 avec recouvrement). L'utilisateur peut aisément visualiser le graphe et interagir dessus. Il dispose de nombreuses opérations interactives (e.g. zoom, sélection, navigation) sur les noeuds/données et finalement obtient une classification en 3(i)).*



(a) initialisation

(b) sélection d'un 1^{er} cluster

(c) ^{er} cluster sélectionné (couleur marron attribuée)

(d) sélection d'un 2^{e} cluster

(e) 2^{e} cluster sélectionné (couleur orange attribuée)

(f) sélection d'un 3^{e} cluster

(g) 3^{e} cluster sélectionné (couleur bleu attribuée)

(h) une classification est obtenue de manière interactive

FIG. 4 – *Exemple complet d'une classification visuelle et interactive en 2D sur un graphe construit avec notre méthode à partir de la base de données WINE (178 données, 3 classes). L'utilisateur peut aisément visualiser le graphe et interagir dessus. Il dispose de nombreuses opérations interactives (i.e. zooms, sélection) sur les noeuds/données et finalement obtient une classification en 4(h)).*