Université Lumière Lyon2
Année 2003

Thèse
pour obtenir le grade de
Docteur
en
Informatique

présentée et soutenue publiquement par

Pierre-Emmanuel JOUVE
le 10 décembre 2003

Apprentissage Non Supervisé et Extraction de Connaissances à partir de Données

préparée au sein du laboratoire ERIC
Equipe de Recherche en Ingénierie des Connaissances

sous la direction de
Nicolas Nicoloyannis

devant le jury, composé de:

Jean-Paul Rasson, Rapporteur	Professeur, Facultés Universitaires N.D. de la Paix, Namur
Gilles Venturini, Rapporteur	Professeur, Université de Tours
Mohand-Saïd Hacid, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Michel Lamure, Examinateur	Professeur, Université Claude Bernard-Lyon 1
Gilbert Ritschard, Examinateur	Professeur, Université de Genève
Nicolas Nicoloyannis, Directeur de thèse	Professeur, Université Lumière-Lyon 2

Table des matières

7 Conclusion

"On ne fait jamais attention à ce qui a été fait ; on ne voit que ce qui reste à faire."

- Marie Curie -

Les travaux présentés dans cette thèse participent donc en premier lieu de l'intégration de la classification non supervisée au sein d'un processus d'ECD.

Nous nous sommes penchés sur différentes étapes de ce processus (la sélection de variables au chapitre 5, l'application d'algorithmes de classification non supervisée au chapitre 3, l'évaluation de la validité des résultats au chapitre 4) en nous efforçant de proposer de nouvelles solutions originales permettant d'une part la résolution de problèmes ou l'intégration d'exigences issues de la pratique et d'autre part de bien prendre en compte le rôle central de l'utilisateur dans le processus ECD.

Ainsi, la méthode de sélection de variables proposée (pour la classification non supervisée) semble permettre le traitement de jeux de données plus volumineux en limitant le temps de traitement tout en assurant un bon niveau de qualité des résultats. Cette contribution nous apparaît intéressante d'une part car peu de méthodes permettant l'automatisation de ce processus existent, et d'autre part car certaines des méthodes existantes (telles celles fondées sur l'analyse factorielle) nécessitent d'appliquer l'algorithme de classification non supervisée sur un espace de représentation des données modifié (les axes principaux de l'analyse factorielle par exemple), rendant ainsi plus complexe l'interprétation des résultats.

La mise au point de la méthode de classification non supervisée (pour données catégorielles) KEROUAC a, quant à elle, été motivée par le désir de disposer d'une méthode à l'utilisabilité augmentée : sa mise en œuvre ne nécessite pas une expertise particulière tout comme l'analyse de ses résultats ne suppose pas une interprétation complexe à mener. De plus, cette méthode autorise le traitement de données manquantes ou encore l'intégration de contraintes, de connaissances qui constituent autant d'éléments bien utiles en pratique. Enfin, les éléments classiquement définis comme essentiels à une bonne méthode de classification non supervisée (validité et stabilité des résultats fournis, coût calculatoire "allégé"...) constituent autant de points non négligés par la méthode KEROUAC.

Evaluer aisément la qualité/validité des résultats d'un processus de classification non supervisée correspond là encore à une préoccupation importante tant d'un point de vue pratique qu'académique. Aussi avons nous proposé une méthodologie originale d'évaluation/comparaison de la validité de classification non supervisée, méthodologie s'appuyant sur une analyse graphique relativement intuitive. Cette méthodologie semble de plus se démarquer de la majorité des approches existantes de part son coût calculatoire limité combiné à la possibilité de comparer toutes formes de partitions, et de part l'aspect caractérisation visuelle d'un jeu de données qu'elle propose.

Enfin, de manière à intégrer mieux encore des exigences résultant de la pratique (traitement de données distribuées physiquement, traitement de données hétérogènes, réutilisation de connaissances, nécessité d'accélérer le processus de classification non supervisée...) nous avons introduit la problématique "Cluster Ensembles" et proposé l'utilisation de trois méthodes pour sa résolution. Les intérêts principaux de cette partie sont de souligner de manière simultanée les différents apports de l'agrégation de classification non supervisée et de proposer d'utiliser deux méthodes de classification non supervisée efficientes pour procéder à l'agrégation. Ce dernier point étant particulièrement intéressant car il semble démontrer que la simple réutilisation de méthodes existantes permet d'atteindre des résultats qualitativement aussi intéressants que ceux obtenus avec des méthodes spécifiques (en permettant même une accélération du processus d'agrégation pour la méthode K-Modes, ou la limitation du paramétrage dans le cas de la méthode KEROUAC).

Dénominateurs communs à l'ensemble de ces travaux, les concepts de comparaisons par paires, d'agrégation de préférences (concepts sous-jacents au critère de Condorcet) se sont avérés être des outils puissants.

Outre la présentation de solutions concrètes évoquées plus tôt, cette thèse vise également à indiquer que ces concepts véhiculent certainement des éléments de solutions à de nombreux problèmes alors qu'ils apparaissent actuellement sous-exploités.

Assistés des éléments théoriques fournit par la S-Théorie de Michaud [Mic87], [Mic91] (éléments visant notamment à la résolution efficace de problèmes d'optimisation posés dans le cadre précis de l'agrégation de préférences) nous pensons que ces concepts peuvent constituer des bases alternatives pour la mise au point de futures méthodes et méthodologies efficientes dans le cadre de l'ECD.

Des travaux additionnels, sortant du domaine du "Non Supervisé" et touchant au domaine du "Supervisé" ont également été abordés, soit en tant que

travaux aboutis (la méthode de sélection de variables pour l'apprentissage supervisé du chapitre 4), soit sous forme de prototype (la méthode d'apprentissage supervisé par apprentissage non supervisé sous contraintes du chapitre 3), ou simplement évoqués comme travaux à venir (mise au point de méthode d'apprentissage généralisé par le biais de KEROUAC (voir chapitre 3), utilisation des méthodes d'agrégation de classifications non supervisées dans le cadre de l'agrégation de classifieurs supervisés (voir chapitre 6)).

Concernant la méthode de sélection de variables pour l'apprentissage supervisé, les expérimentations menées la place au niveau des méthodes de référence actuelles tant pour la qualité des modèles d'apprentissage qu'elle permet de bâtir, que pour la réduction de l'espace de représentation des données qu'elle implique et le temps d'exécution qui lui est associé. Enfin, nous avons vu que cette méthode était fondée sur un certain nombre d'hypothèses dont certaines apparaissent trop strictes et dont la relaxation (envisagée dans des travaux futures) devrait (peut être ?) permettre l'amélioration de ces performances.

La méthode d'apprentissage non supervisé sous contraintes (qui n'est actuellement qu'un prototype) semble, quant à elle, fournir des modèles de qualité relativement similaire à celle des approches classiques, et son développement devrait permettre de répondre à un certain nombre de questions intéressantes touchant notamment à l'intérêt de l'apprentissage semi-supervisé.

Ces travaux (participant non plus de l'intégration de la classification non supervisée au sein du processus d'ECD mais de l'utilisation de la classification non supervisée pour d'autres éléments constitutifs d'un processus d'ECD) présentent donc un intérêt en tant que tel, mais, leur intégration dans ce document est également motivée par la volonté de souligner l'importance (et peut être la prééminence ?) de la structuration de l'information pour le processus d'ECD (la structuration de l'information étant abordée ici par le biais de la classification non supervisée).

En effet, si la relation entre classification non supervisée et apprentissage supervisé par les méthodes de type plus proche voisin est évidente (le concept "qui se ressemblent s'assemblent" sous-jacent à ce type de méthode d'apprentissage supervisé régit également la classification non supervisée) les méthodes d'apprentissage semi-supervisé mettant à profit la structure interne de l'information dans le cadre d'apprentissage supervisé sont encore peu nombreuses et peu exploitées. Ce constat peut être généralisé à l'ensemble des méthodes d'ECD : un faible nombre d'entre elles exploitent vraiment la structuration de l'information.

Le prototype de méthode d'apprentissage non supervisé sous contraintes, la méthode de sélection de variables pour l'apprentissage supervisé (présentée comme une méthode visant à déterminer l'espace de représentation des données tel que sa structure interne reflète au mieux la réalité à apprendre) doivent ainsi être, en partie, considérés comme autant d'éléments contribuant à montrer l'intérêt que peut revêtir l'apport de la structuration dans le cadre

d'un processus d'ECD. Notons enfin que, à l'opposé, il nous semble parfois utile d'adopter une perspective apprentissage supervisé afin de "féconder" des travaux touchant à la classification non supervisée comme le montre la méthode de sélection de variables pour l'apprentissage non supervisé qui peut être envisagée comme une généralisation de la méthode de sélection de variables proposée dans le cadre supervisé (plus spécifiquement une méthode de sélection de variables dans un cadre supervisé où plusieurs variables endogènes cohabitent).

Nous synthétisons finalement les contributions de cette thèse, les relations unissant les univers "Supervisé" et "Non Supervisé" apparaissant dans cette dissertations, ainsi que les travaux à venir sur les figures 7.1, 7.2.

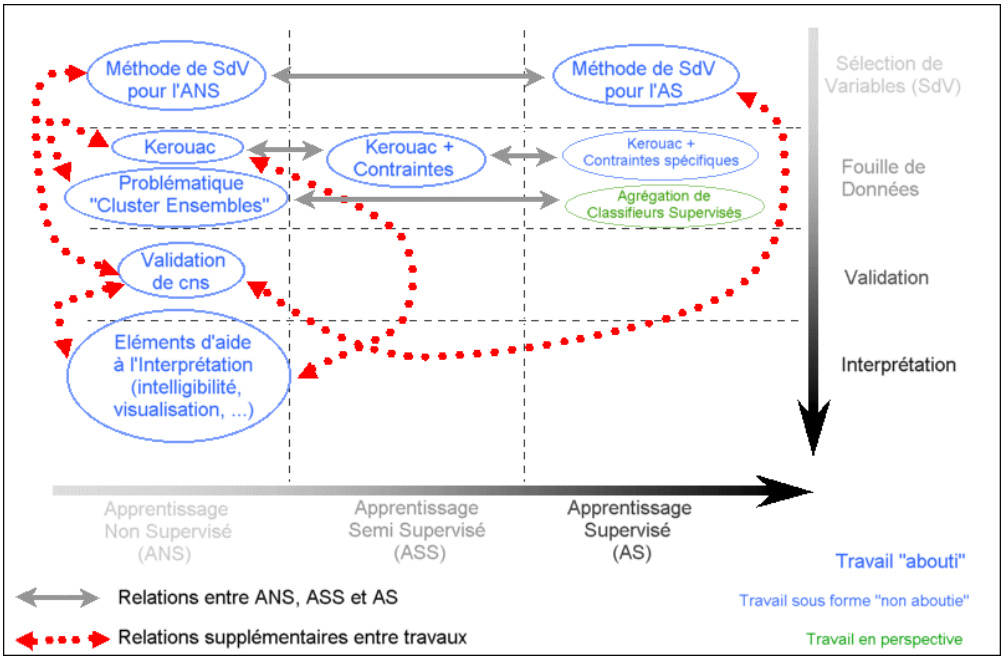

FIG. 7.1 –: *Synthèse des contributions, relations entre contributions et relations unissant entre "Supervisé", "Semi-Supervisé" et "Non Supervisé"*

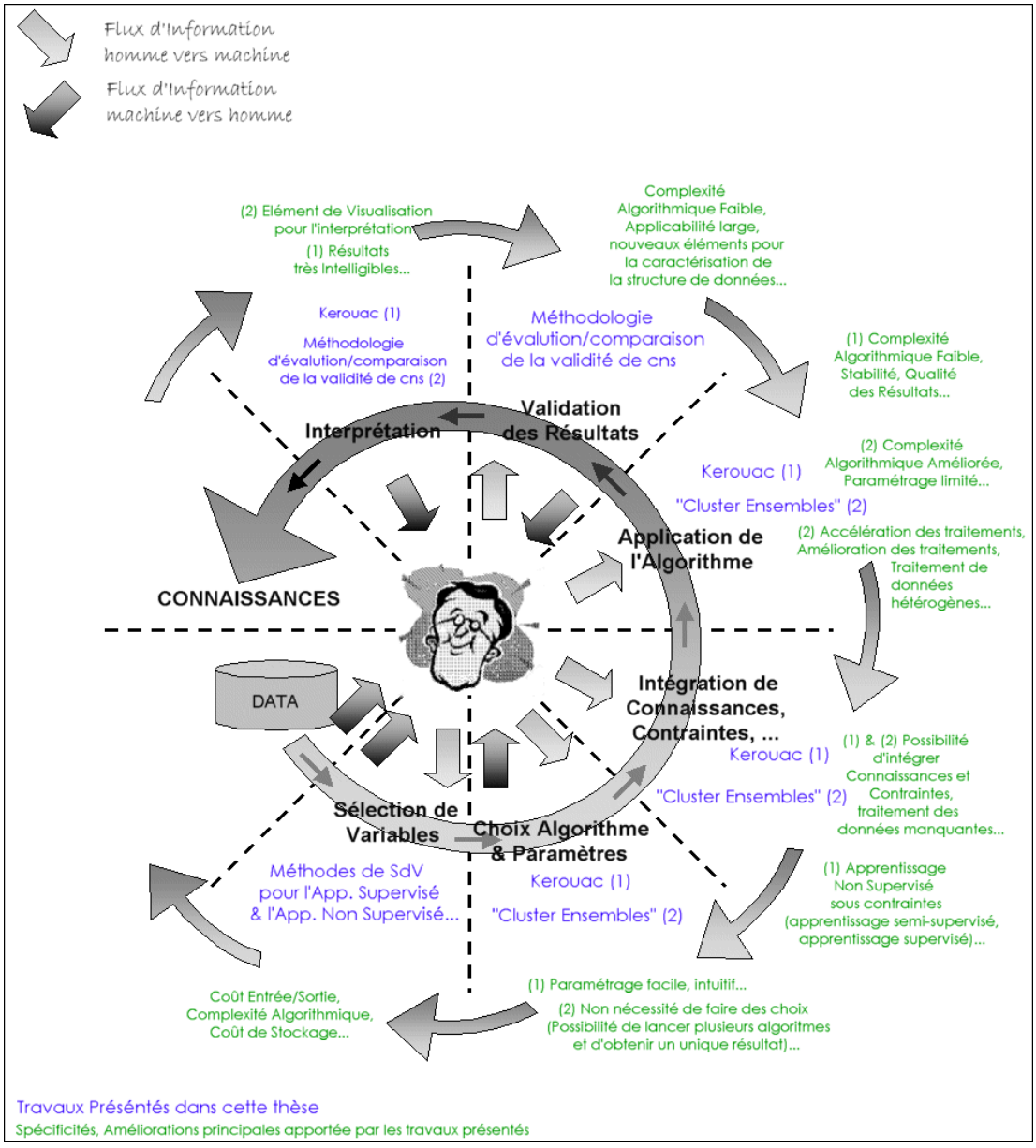

FIG. 7.2 –: *Synthèse des contributions*