Clustering dynamique d'un flot de données : un algorithme incrémental et optimal de détection des maxima de densité

Alain Lelu

LASELDI / Université de Franche-Comté
30 rue Mégevand – 25030 Besançon cedex
alain.lelu@univ-fcomte.fr

Résumé. L'extraction non supervisée et incrémentale de classes sur un flot de données (*data stream clustering*) est un domaine en pleine expansion. La plupart des approches visent l'efficacité informatique. La nôtre, bien que se prêtant à un passage à l'échelle en mode distribué, relève d'une problématique *qualitative*, applicable en particulier au domaine de la veille informationnelle : faire apparaître les évolutions fines, les « signaux faibles », à partir des thématiques extraites d'un flot de documents. Notre méthode GERMEN localise de façon exhaustive les maxima du paysage de densité des données à l'instant t, en identifiant les perturbations locales du paysage à t-1 et modifications de frontières induites par le document présenté. Son caractère optimal provient de son exhaustivité (à une valeur du paramètre de localité correspond un ensemble unique de maxima, et un découpage unique des classes qui la rend indépendante de tout paramètre d'initialisation et de l'ordre des données.

1 Introduction et objectifs

Pour rendre compte avec exactitude des évolutions temporelles, cruciales dans beaucoup de domaines d'application (ex. : veille d'information), il est nécessaire à notre avis :
1) de partir d'une base stable, c'est-à-dire d'une classification :
- indépendante de l'ordre de présentation des données (exigence n°1),
- indépendante des conditions initiales, que ce soit d'un choix de « graines de classes » arbitraires ou dépendantes des données (exigence n°2),
- impliquant un minimum de paramètres, un seul si possible, pour réduire l'espace des choix et tendre vers un maximum de vérifiabilité et de reproductibilité (exigence n°3).
2) d'ajouter aux contraintes d'une bonne classification celle de l'incrémentalité (exigence N°4), afin de saisir les évolutions au fil de l'eau : rectifications de frontières entre classes, apparition de nouvelles classes, voire de « signaux faibles »... Pour nous, il y a incrémentalité véritable si le résultat de la classification est indépendant de l'ordre des données présentées antérieurement (exigence N°5), tout en découlant des données antérieures, par un historique pouvant faire l'objet d'interprétations.

Notre démarche a été de concevoir une méthode où la contrainte d'incrémentalité participer d'un tout cohérent, en vue d'aboutir à tout instant à une classification qui ait du sens, et dont la différence de représentation par rapport à l'instant précédent ne provient que des

effets du temps, et non du mélange de ceux-ci avec la variabilité propre de l'algorithme, à la différence des principales méthodes de classification non supervisée.

2 Etat de l'art

Les *méthodes hiérarchiques*, divisives ou agglomératives, souvent conviviales et efficaces, satisfont à nos exigences 1 à 3 d'unicité des résultats. Mais au regard de la qualité des partitions obtenues à un niveau donné de l'arbre, un consensus existe pour leur préférer les méthodes à centres mobiles : un modèle hiérarchique de partitions emboîtées impose des déformations à une réalité ayant toutes chances de se rapprocher d'une organisation en treillis de partitions (par ex. treillis de Galois, pour des descripteurs binaires).

Les méthodes procédant par *agrégation autour de centres mobiles*, comme les K-means et leurs nombreuses variantes, font partie d'une famille basée sur l'optimisation d'un indicateur numérique global de qualité de la partition, dans laquelle prennent place les méthodes utilisant la procédure EM (*Expectation Maximization*) – cf. Buntine (2002). Ce problème d'optimisation étant NP-difficile, on ne sait que les faire converger vers un optimum local qui dépend de leur initialisation (par ex. positions initiales des centres choisies arbitrairement, ou en fonction des données), voire de l'ordre des données. Ce qui les disqualifie vis-à-vis de notre exigence N°2 en théorie comme en pratique. Un bon nombre de variantes incrémentales de ces méthodes ont été proposées, dont on trouvera une revue partielle dans Gaber et al. (2005). Beaucoup sont issues de l'action DARPA « Topic Detection and Tracking » Mais toutes se concentrent sur l'efficacité informatique pour traiter des flux de dizaines ou centaines de milliers de dépêches d'agences, ou autres documents. C'est aussi dans ce cadre d'efficacité que Simovici et al. (2005) proposent un algorithme glouton pour optimiser un critère de qualité de partition original propre aux descriptions par variables nominales.

A notre connaissance, seules les *méthodes basées sur la densité* satisfont à notre exigence d'unicité des résultats, hors de portée des méthodes ci-dessus. Elles s'appuient sur la notion de densité d'un nuage de points, locale par définition : étant donné 1) un nuage de points multidimensionnel, 2) une définition de la densité en chaque point de cet espace, 3) la valeur du paramètre de localité de cette fonction densité (son « rayon »), le paysage de densité qui en découle est unique et parfaitement défini. L'énumération de l'ensemble de ses pics – ou éventuels plateaux - représente un optimum absolu ; ces pics balisent des noyaux homogènes de points ; dans les zones intermédiaires, on peut définir de diverses façons des zones d'influence des noyaux. Trémolières (1994) a proposé un algorithme général, dit de percolation, indépendant de la définition de la densité et du type de données, pour délimiter rigoureusement les noyaux, les points-frontière ambivalents et les points atypiques. Il procède par baisse progressive du niveau de densité depuis le point le plus dense, et diffusion autour des noyaux qui apparaissent successivement. D'autres travaux retrouvent le même principe de repérage des noyaux denses, le plus souvent avec une définition spécifique de la densité, et d'extensions de diverses sortes à partir des noyaux : Moody (2001), Guénoche (2004), Batagelj (2002)... A noter que ces méthodes peuvent se traduire en termes de partitionnement de graphe : définir une densité implique d'avoir fixé des relations de voisinage, donc un graphe. DBSCAN d'Ester et al. (1996) utilise une définition de la densité au moyen de deux paramètres, dont l'un fixe le seuil à partir duquel les noyaux sont constitués et étendus. Ertöz et al. (2003) utilisent la notion de K plus proches voisins, qui permet de définir un « rayon adaptatif » ; plusieurs possibilités existent alors pour définir une densité au sein de ce voisinage.

Ici aussi des méthodes incrémentales ont été proposées : ainsi une version incrémentale de DBSCAN d'Ester et al. (1998), ou la méthode de Gao et al. (2005), pour des descripteurs quantitatifs (âge, revenus, …). Mais c'est dans le domaine des protocoles auto-organisateurs de réseaux de communication radio dits « ad hoc » que le thème de l'incrémentalité pour le partitionnement dynamique de graphes évolutifs a été principalement abordé - ex. : Mitton, Fleury (2003) – avec des objectifs d'application assez différents des nôtres : les voisinages se modifient à chaque pas de temps, sans nécessairement comporter d'entrées ou sorties du réseau, l'optimalité est moins recherchée qu'une stabilité relative de la composition des classes (strictes) et de l'identité des chefs de classe (*clusterheads*). Le principe-clé est celui d'un algorithme intégralement *distribué*, pour lequel la connaissance à l'instant t par chaque unité à classer de ses voisins et des voisins de ses voisins (2-voisinage) est suffisante.

3 Algorithme incrémental GERMEN : modifications *locales* des voisinages, des densités et des classes

Nous avons publié et expérimenté sur des données documentaires [cf. Lelu, François (2003)] un algorithme de percolation modifié, aboutissant à extraire 1) les points isolés, 2) les noyaux stricts exclusifs d'une classe, 3) les points ambivalents appartenant à plusieurs classes, tous ces points se projetant à des hauteurs diverses sur chaque axe de classe. Nous y renvoyons en ce qui concerne la transformation préalable des données.

A la différence de cet algorithme, celui que nous présentons ici est incrémental. Dans cette optique, nous nous donnons l'état d'un graphe de similarité des K plus proches voisins à l'instant t, valué et orienté, dont les nœuds ont été caractérisés par leur densité, ainsi que par leur « couleur », c'est-à-dire par leur rattachement à un éventuel nœud « chef de classe », dont le numéro constitue l'étiquette. L'arrivée d'un nouveau nœud va perturber localement cet état : un certain nombre de nœuds dans le voisinage direct ou indirect du nouvel arrivant vont voir leur densité changer – ce changement du paysage de densité induisant à son tour un réajustement des zones d'influence des chefs de classe, voire des changements de chefs de classe, ou l'apparition de nouvelles classes (cf. plus bas le pseudocode détaillé).

Plusieurs règles d'héritage du numéro du (ou des) chef(s) de classe sont possibles. Mais dans tous les cas nous avons affaire à la mise à jour des « couleurs » d'un paysage de densité sous l'effet 1) de l'arrivée d'un nouveau point (changement structurel), 2) d'un changement localisé des densités (changement quantitatif). Si la mise à jour de la couleur d'un point ne dépend que de la couleur de ses voisins « surplombants », de densité supérieure, alors à paysage de densité donné et à graphe de voisinage donné on obtiendra un coloriage unique : l'attribution des classes sera elle aussi indépendante de l'ordre des données

On construit progressivement et en la mettant constamment à jour une structure de données comportant pour chaque noeud la liste de ses &- et 2-voisins, sa densité, et son (ou ses) numéros de chef(s) de classe. A chaque arrivée d'un vecteur (nœud) nouveau, on calcule les changements de densité induits dans son 2-voisinage (il ne peut pas y en avoir ailleurs du fait de notre définition de la densité comme somme des liens présents dans l'ensemble de chaque 1-voisinage), puis les changements de chef(s) de classe induits. Pour ce faire, on met à jour et on parcourt itérativement la liste des nœuds susceptibles de changer de classe, compte tenu de la règle choisie pour l'extension des classes. Au pire cette liste peut comporter tous les documents antérieurs, mais elle ne peut que se vider (en pratique elle se stabilise autour de

deux ou trois centaines d'éléments en moyenne quand le nombre d'éléments à classer dépasse un millier) :

```
. Initialisation : le premier noeud de la séquence n'a aucun lien, a une densité nulle et son propre numéro comme chef
     de classe.
LCC = Ø  // LCC est la liste des listes de chef(s) de classe pour chaque noeud //

. POUR chaque nouveau noeud :

     // changements de densité induits : //
     .calcul de ses 1- et 2-voisinages entrants et sortants, et des modifications de voisinages occasionnée par son ar-
     rivée ; d'où la liste LL des noeuds touchés par une création / suppr. /modif. de lien.
     .calcul de sa densité à partir de son 1-voisinage.
     .POUR tout noeud de LL, et tout noeud de son 1-voisinage :
          - nouveau calcul de sa densité
     fin POUR

     // changements de chef(s) de classe induits : //
     L = LL
     TANT QUE  la liste L des noeuds susceptibles de changer d'état est non-vide :
          liste LS = Ø
          POUR chaque noeud de L trié par densité décroissante :
               ~ appliquer la règle de changement de classe en fonction des classes des voisins entrants surplombants
               (données par LCC) et de leurs densités.
               ~ si changement :
                    .mettre à jour LCC pour le noeud courant
                    .déterminer les surplombés éventuels (dont le noeud courant est surplombant) ; .incrémenter la
                    liste LS des surplombés.
          fin POUR
          L=LS
     fin TANT QUE
```

Règle d'héritage du (ou des) chef(s) de classe choisie : tout noeud hérite du ou des N° de chef de classe de *tous* ses voisins surplombants, s'il y en a (possibilité d'appartenance à plusieurs classes, polysémie des mots et des textes oblige) ; sinon, on crée une nouvelle classe.

Valeurs ex-aequo de similarité et densité : leur prise en compte est indispensable pour assurer l'indépendance par rapport à l'ordre des données. Les K plus proches voisins d'un noeud peuvent donc être en nombre supérieur à K… En cas de voisins de même densité et dominant leurs voisinages, le numéro du nœud le plus ancien est attribué.

3.1 Illustration du contenu des classes par leurs descripteurs saillants.

Nous définissons comme suit la contribution relative $C_i(k)$ du descripteur i à la classe k, (dont on nomme liens(k) l'ensemble des liens internes) : si x_{it} est le nombre d'occurrences du descripteur i dans le document N°t, de somme $x_{.t}$ pour l'ensemble de ces descripteurs, si $c_i(t,t')$ est la contribution pondérée du descripteur i au lien entre les documents t et t', de densités respectives d(t) et d(t'), alors :

$c_i(t,t') = [d(t)\ d(t')]^{½}\ [x_{it}\ /x_{.t}]^{½}\ [x_{it'}\ /x_{.t'}]^{½}$ et $C_i(k) = \Sigma_{liens(k)}\ c_i(t,t') / \Sigma_k\ \Sigma_{liens(k)}\ c_i(t,t')$

3.2 Complexité informatique

Dans son implantation actuelle, l'introduction incrémentale du $n^{ème}$ document coûte de l'ordre de O(n) en temps de calcul, donc l'algorithme complet est en $O(n^2)$. Des optimisations sont possibles dans le calcul des similarités et des K plus proches voisins (vecteurs-

données très creux) ; mais c'est surtout en l'exécutant en mode distribué, ce à quoi il est adapté par construction, qu'il pourrait être rendu peu dépendant de la taille des données.

4 Exemple d'application, évaluation.

Ci-dessous un exemple de thème obtenu sur une base textuelle de test (193 résumés de la collection Gallimard-Jeunesse, vocabulaire de 888 mots retenus, K=3 plus proches voisins), parmi 28 classes obtenues : à t=193 ; environ 45% des documents appartiennent à des noyaux, 5% sont des isolés, et 50% des ambivalents. Pour évaluer la qualité des résultats, les corpus de test des campagnes TDT citées plus haut sont inadéquats pour deux raisons : 1) le nombre de documents va de 16 000 à plusieurs centaines de milliers, ce qui est pour l'instant un ou deux ordres de grandeur au dessus des possibilités de la version actuelle, 2) ce qui est évalué n'est pas « pur », c'est une chaîne indexation + algorithmes utilisant cette indexation. C'est pourquoi nous avons entrepris – cf. Lelu et al. (2005) - de constituer la base de test « 10 ans de géotechnique dans PASCAL » indexée par l'INIST (en français et en anglais) et étiquetée semi-automatiquement par sujets, dont nous avons analysé l'année 2003 à titre de référence (1200 documents), et qui sera mise à disposition quand l'étiquetage sera terminé.

```
Noyaux :
1  162  3.5463867  Corot,.la.Mémoire.du.paysage.............
1  72   2.5157569  Miró.le.peintre.aux.étoiles.............
1  24   2.1824623  Monet...................................
1  63   1.9537317  Renoir,."Il.faut.embellir"..............
1  118  1.9195755  Poussin,."Je.n'ai.rien.négligé".........
1  89   1.4746671  Manet.:."J'ai.fait.ce.que.j'ai.vu"......
1  139  1.4316663  Le.Caravage,.peintre.et.assassin........
1  64    .5066200  Vuillard,.le.temps.détourné.............
1  45   1.0033725  Géricault,.l'invention.du.réel..........
1  126  1.1061365  Chagall,.ivre.d'images..................
1  164  1.3935725  Blaise.Cendrars,.l'or.d'un.poète........
Documents bivalents :
2  25    .4496930  Toulouse-Lautrec,.les.lumières.de.la.nui
2  156   .4801046  Courbet,.le.poème.de.la.nature..........
2  172  1.3933956  Bacon,.monstre.de.peinture..............
2  178  1.2192515  Léonard.de.Vinci,.art.et.science.de.l'un
2  174  1.03453    Calder,.le.scupture.en.mouvement........
2  52    .7916118  Matisse,.une.splendeur.inouïe...........
Documents trivalents et plus :
3  138   .2398822  Chère.Madame.de.Sévigné.................
3  4     .5631906  Les.surréalistes,.une.génération.entre.l
3  117   .5278936  Le.Cheval,.force.de.l'homme.............
3  125   .3197081  La.Fontaine.ou.les.métamorphoses.d'Orphé
```

```
Termes illustratifs :
144  peintre...........
68   Paris.............
31   monet.............
30   travail...........
30   peinture..........
27   rome..............
22   mouvement.........
22   atelier...........
21   voyage............
21   paysage...........
21   portrait..........
21   france............
20   cheval............
20   lumière...........
18   sculpture.........
17   artiste...........
15   couleur...........
14   famille...........
13   tableau...........
13   guerre............
12   salon.............
12   poésie............
12   chef..............
```

FIG. 1 – *Exemple de thème extrait par GERMEN*

5 Conclusion et perspectives

Nous avons présenté un algorithme de classification non supervisée répondant aux exigences d'un suivi dynamique rigoureux d'un flux de documents : il est à la fois optimal, au sens de la description exhaustive d'un paysage de densité adaptative, et incrémental. Son efficacité actuelle est suffisante pour traiter un flux de plusieurs milliers de documents. Son passage à l'échelle est possible en mode distribué, par construction.. Mais il reste surtout à explorer en grandeur réelle et évaluer les diverses applications et usages nouveaux possibles, avec les problèmes qui vont avec : efficacité informatique, ergonomie de représentation, et

définition même de ce que peut être la représentation dynamique et interactive d'une réalité évolutive - problème peu abordé jusqu'à présent.

Références

Batagelj, V., M. Zaversnik, (2002) An o(m) algorithm for cores decomposition of networks, University of Ljubljana, *preprint series* Vol. 40, 799,

Buntine W. L. (2002). Variational Extensions to EM and Multinomial PCA. *ECML 2002*

Ertöz L., M. Steinbach, V.Kumar (2003). Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data.. *SIAM /*SDM '03).

Ester M., H.-P. Kriegel, J. Sander, M. Wimmer, X. Xu (1998). Incremental Clustering for Mining in a Data Warehousing Environment. *VLDB 1998*: 323-333

Gaber M., A. Zaslavsky and S. Krishnaswamy (2005). Mining Data Streams: A Review. *SIGMOD Record*, 34(2).

Gao J., J. Li, Z. Zhang, P.-N. Tan (2005). An Incremental Data Stream Clustering Algorithm Based on Dense Units Detection. *Proc. PAKDD'05*, pp. 420-425.

Guénoche A. (2004). Clustering by vertex density in a Graph. *Meeting of the IFCS*. Chicago, Classification, Clustering and Data Mining, D. Banks et al. (Eds.), Springer, 15-23.

Lelu A, P. Cuxac, J. Johansson. (2005), Classification dynamique d'un flux documentaire : évaluation statique préalable. Soumis à. *JADT 2006*, L. Lebart, A. Salem, org..

Lelu A., C. François (2003), Un algorithme de détection de maxima de densité basé sur la distance distributionnelle : application à la classification optimale fine d'un corpus documentaire, *10èmes Rencontres SFC*, Dodge Y., Melfi G. eds., Pr. Acad. Neufchâtel.

Mitton N., A. Busson, E. Fleury (2004) Self-organization in large scale ad hoc networks. *The Third Annual Mediterranean Ad Hoc Networking Workshop*, Bodrum, Turkey.

Moody J. (2001) – Identifying dense clusters in large networks, *Social Networks*, vol. 23.

Simovici D., N. Singla, M. Kuperberg (2004) Metric Incremental Clustering of Nominal Data. *Proceedings of ICDM 2004*, Brighton, UK, pp. 523-527

Trémolières R.C. (1994). Percolation and multimodal data structuring - *New Approaches in Classification and Data Analysis*, Diday E. et al. (eds.), 263-268, Springer Verlag, Berlin.

Summary

Data stream clustering is an ever-expanding subdomain of knowledge extraction. Most of the past and present research effort aims at efficient scaling up for the huge data repositories. Our approach focuses on qualitative improvement on medium-sized databases, mainly for "weak signals" detection and precise tracking of topical evolutions in the framework of information watch – though scalability is intrinsically guaranteed in a possibly distributed implementation. Our GERMEN algorithm exhaustively picks up the whole set of density peaks of the data at time t, by identifying the *local* perturbations induced by the current document vector, such as changing cluster borders, or new/vanishing clusters. Optimality, data-rank independence, yields from the uniqueness 1) of the density landscape for any value of our zoom parameter, 2) of the cluster allocation operated by our border propagation rule.