Bordures statistiques pour la fouille incrémentale de données dans les Data Streams

Jean-Emile Symphor*, Pierre-Alain Laur*

*GRIMAAG-Dépt Scientifique Interfacultaire,
Université des Antilles et de la Guyane, Campus de Schoelcher,
B.P. 7209, 97275 Schoelcher Cedex, Martinique, France
{je.symphor,palaur}@martinique.univ-ag.fr.

Résumé. Récemment la communauté Extraction de Connaissances s'est intéressée à de nouveaux modèles où les données arrivent séquentiellement sous la forme d'un flot rapide et continu, $i.e.$ les data streams. L'une des particularités importantes de ces flots est que seule une quantité d'information partielle est disponible au cours du temps. Ainsi après différentes mises à jour successives, il devient indispensable de considérer l'incertitude inhérente à l'information retenue. Dans cet article, nous introduisons une nouvelle approche statistique en biaisant les valeurs supports pour les motifs fréquents. Cette dernière a l'avantage de maximiser l'un des deux paramètres (précision ou rappel) déterminés par l'utilisateur tout en limitant la dégradation sur le paramètre non choisi. Pour cela, nous définissons les notions de bordures statistiques. Celles-ci constituent les ensembles de motifs candidats qui s'avèrent très pertinents à utiliser dans le cas de la mise à jour incrémentale des streams. Les différentes expérimentations effectuées dans le cadre de recherche de motifs séquentiels ont montré l'intérêt de l'approche et le potentiel des techniques utilisées.

1 Introduction

Ces dix dernières années un grand nombre de travaux ont été proposés pour rechercher des motifs fréquents dans de grandes bases de données. En fonction des domaines d'applications les motifs extraits sont soit des itemsets (Srikant, 1995; Zaki, 2001; Pei et al., 2001; Ayres et al., 2002) soit des séquences (Agrawal et al., 1993; Han et al., 2000). Récemment les travaux issus de la communauté des chercheurs en base de données et en fouille de données considèrent le cas des data streams où l'acquisition des données s'effectue de façon régulière, continue ou incrémentalement et cela sur une durée longue voire éventuellement illimitée.

Compte tenu de la grande quantité d'information mise en jeu dans le cas des data streams, le problème de l'extraction de motifs fréquents est toujours d'actualité ((Li et al., 2004; Jin et al., 2003; Demaine et al., 2002; Manku et Motwani, 2002; Golab et Ozsu, 2003; Karp et al., 2003)). Dans ce contexte, un motif est dit *θ-fréquent* s'il est observé au moins une fraction θ, appelée support du motif, sur tout le stream. Le paramètre $theta$, tel que $0 < \theta < 1$, est fixé par l'utilisateur.

Dans le cas des data streams, sujets à des mises à jour régulières et fréquentes, les approches traditionnelles ne conviennent pas car les résultats obtenus pour l'ancienne base ne sont que partiellement valables pour la nouvelle et il n'est pas envisageable de relancer l'algorithme sur la base de données mise à jour. En effet, dans tous les travaux développant une approche par mise à jour incrémentale (Masseglia et al., 2003; Cheng et al., 2004), la problématique principale d'optimisation et de performance consiste à construire et à maintenir, au fur et à mesure des différentes mises à jour successives, un ensemble de motifs candidats. Celui-ci est utilisé pour mettre à jour les motifs fréquents et éviter de relancer l'algorithme depuis zéro.

Il convient également de souligner une autre caractéristique intrinsèque des data streams qui découle du fait que la connaissance du stream n'est que partielle quel que soit l'instant considéré. En conséquence, il est nécessaire de prendre en compte l'incertitude engendrée par la connaissance toujours incomplète du data stream. Précisément, cela se traduit dans le cas de la recherche de motifs fréquents en soulignant que les motifs fréquents obtenus ne sont en fait que des motifs fréquents observés. En fait, à cause de cette incertitude, deux sources d'erreurs doivent être considérées :

- Les motifs observés comme fréquents ne sont peut être plus du tout fréquents sur une longue période d'observation du stream.
- Inversement des motifs classés comme non fréquents peuvent le devenir sur une plus longue période d'observation du stream.

Pour éviter ces erreurs, il est nécessaire de développer une approche pour connaître si un motif est fréquent sur une partie déjà examinée du stream. Cette approche doit de plus être prédictive pour savoir avec quelle probabilité un motif est fréquent ou non sur l'ensemble du stream. De nombreuses applications, par exemple dans le domaine de la prévision météorologique ou encore dans l'analyse de tendance en finance nécessitent ce type d'approche. Même en disposant d'une très grande partie du stream, d'un point de vue statistique, il est impossible de s'affranchir de ces deux sources d'erreur (Vapnik, 1998). Notre objectif sera donc d'essayer d'approcher le mieux possible une solution optimale.

Dans cet article nous proposons une approche qui permet, tout en considérant l'incertitude inhérente à la connaissance des streams, de construire et de maintenir des ensembles de motifs candidats bien choisis. Ceci constitue un préalable fondamental et nécessaire à toute approche pertinente dans le cadre de la mise à jour incrémentale des data streams.

La suite de la présentation est organisée de la façon suivante. Dans le paragraphe 2, nous introduisons les concepts qui permettent de contrôler l'incertitude découlant des sources d'erreurs. Au paragraphe 3, nous montrons comment obtenir les ensembles de motifs pertinents pour effectuer la mise à jour incrémentale. Le paragraphe 4 présente une expérimentation de notre approche, suivie d'une analyse comparative avec des travaux connexes au paragraphe 5. Nous concluerons notre étude au paragraphe 6.

2 Supports statistiques

2.1 Définitions

Pour formaliser notre problématique, nous définissons les différents ensembles utilisés. Le data stream est obtenu à partir d'un échantillonnage effectué sur un domaine X potentiellement très grand qui contient tous les motifs possibles (figure 1). Chaque motif est échantillonné in-

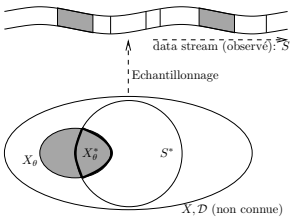

FIG. 1 – *Le problème.*

dépendemment à partir d'une distribution $\mathcal{D}$ sur laquelle nous ne formulons aucune hypothèse, excepté celle de dire qu'il n'y a pas de biais. Nous renvoyons le lecteur intéressé par des approches prenant en compte le biais dans les cas de fouilles de données supervisées aux travaux de (Fan et al., 2004; Wang et al., 2003). La partie du stream observée est représentée par S dans la figure 1. A partir d'une valeur fixée par l'utilisateur du paramètre θ, le support théorique, on voudrait connaître l'ensemble des motifs vrais θ-*fréquents* de X. Cet ensemble nommé X_θ est représenté en grisé sur la figure 1. Hormis l'incertitude et les aspects liés à l'estimation statistique, l'approximation de l'ensemble X_θ recouvre un aspect combinatoire qui provient de la très grande taille de X même si cet ensemble peut être fini. Nous nommons S l'ensemble des motifs observés extraits du stream avec $|S| = m$ ($|S| << |X|$). Nous réduisons cette différence à l'aide d'un algorithme qui permet d'obtenir un super ensemble S^* de S avec $|S^*| = m^* > m$. Typiquement, S^* contient des motifs supplémentaires obtenus à partir d'une généralisation des éléments de S (Mannila et Toivonen, 1997). Ce n'est pas le propos de cet article de traiter l'aspect combinatoire, l'essentiel est que S^* ne sera jamais suffisamment grand pour englober X_θ, au regard de la façon dont il est construit (voir figure 1). Ainsi l'importance du problème d'estimation statistique demeure entier.

Soit l'ensemble X^*_θ (figures 1 et 2) qui représente l'ensemble des motifs vrais θ-*fréquents* de X pour l'ensemble S^*, nous rappelons que cet ensemble ne peut être connu totalement compte tenu non seulement du fait que $|S| << |X|$, mais aussi des deux sources d'erreurs précédemment indiquées qui en découlent. Nous recherchons θ' pour approcher au mieux l'ensemble X^*_θ, en utilisant $S^*_{\theta'}$ (figure 2) qui représente l'ensemble des motifs observés θ'-*fréquents* du stream dans S^*.

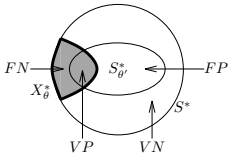

FIG. 2 – *Estimation de l'erreur.*

Pour apprécier l'approximation effectuée, consécutivement aux deux sources d'erreurs , les sous-ensembles X^*_θ et $S^*_{\theta'}$ nous permettent de définir les quatre paramètres suivants (figure 2) :

- VP (vrais positifs) : représente l'ensemble des motifs, vrais θ-*fréquents* de X^*_θ et également observés θ'-*fréquents* de S^*.
- FP (faux positifs) : représente l'ensemble des motifs, vrais non θ-*fréquents* de X^*_θ mais observés θ'-*fréquents* de S^*.
- FN (faux négatifs) : représente l'ensemble des motifs, vrais θ-*fréquents* de X^*_θ mais observés non θ'-*fréquents* de S^*.
- VN (vrais négatifs) : représente l'ensemble des motifs, vrais non θ-*fréquents* du stream et également observés non θ'-*fréquents* de S^*.

A partir des ensembles définis ci-dessus, on peut indiquer les formules de la précision et du rappel qui permettent d'estimer l'approximation effectuée.

$$\mathtt{P} = VP/(VP+FP) \quad (1) \qquad \mathtt{R} = VP/(VP+FN) \quad (2)$$

La précision permet de quantifier la proportion des motifs θ-*fréquents* estimés qui sont en fait non vrais θ-*fréquents*, en dehors de $S^*_{\theta'}$. Si on cherche à Maximiser P, cela revient à minimiser la première source d'erreurs. Symétriquement, le rappel permet de quantifier la proportion de motifs, vrais θ-*fréquents* manquant dans $S^*_{\theta'}$. Si on cherche cette fois à maximiser R, cela revient à minimiser la seconde source d'erreurs.

2.2 Choix de θ'

Définition 1 *θ' est un support statistique pour θ, $0 < \theta' < 1$, s'il est utilisé pour approcher les motifs vrais θ-fréquents du stream.*

Une approche naïve pour approcher au mieux l'ensemble X^*_θ consisterait à choisir $\theta' = \theta$; autrement dit, la question que l'on pourrait se poser est de savoir si θ est un support statistique pour lui-même. Malheureusement, la principale et seule propriété de l'ensemble $S^*_{\theta'}$ dans ce cas, est qu'il tend à correspondre avec une probabilité de 1 à l'ensemble X^*_θ lorsque le cardinal de l'ensemble S^* tend vers ∞ selon le lemme de Borel-Cantelli (Devroye et al., 1996). Cela reviendrait à connaître tout le stream, ce qui n'est pas possible en pratique. Le théorème de Glivenko-Cantelli permet de montrer que l'on peut borner l'erreur commise sur les motifs vrais θ-*fréquents* en fonction de divers paramètres parmi lesquels le cardinal de l'ensemble S^*, mais on ne peut pas faire mieux. Bien souvent, en traitement de l'information ou encore en medecine, plutôt que de simplement borner l'erreur, il est beaucoup plus important d'estimer et de contrôler des parties de l'erreur. On peut par conséquent développer une approche qui consiste à maximiser soit la précision P soit le rappel R. Le choix des valeurs aux limites de θ' pourraient convenir en ce sens où l'on maximise la précision ou le rappel mais ces valeurs sont inintéressantes pour les applications en fouille de données. Par exemple, si on choisit $\theta' = 0$, on obtient $S^*_0 = S^*$ et ainsi $\mathtt{R} = 1$. Mais, nous avons également dans ce cas $\mathtt{P} = |X^*_\theta|/|S^*|$, qui correspond à une valeur trop faible pour bien des applications, et donc tous les éléments de S^* sont considérés comme des motifs vrais θ-fréquents du stream. On pourrait aussi choisir $\theta' = 1$ pour être sûr de maximiser P cette fois, mais il en découle que $\mathtt{R} = 0$ et il se pourrait qu'aucun élément de S^* ne soit un motif vrai θ-fréquent.

Ces exemples avec les valeurs aux limites de θ' nous permettent de bien cadrer les principes de notre approche. L'idée générale est de choisir subtilement θ' différent mais suffisamment

proche de θ, respectivement plus grand ou plus petit que θ, de sorte qu'il soit possible de contrôler en maximisant soit la précision soit le rappel, pour garantir avec une très forte probabilité que $\mathtt{P} = 1$ ou respectivement que $\mathtt{R} = 1$ tout en limitant la dégradation du paramètre non contrôlé. On obtient ainsi un ensemble $S^*_{\theta'}$ pas trop petit contenant des informations significatives. Il y a une barrière statistique autour de θ qui empêche que θ' ne soit ni trop proche ni trop éloigné de θ pour conserver la contrainte que $\mathtt{P} = 1$ ou alors que $\mathtt{R} = 1$, avec une forte probabilité. Notre objectif pour maximiser la précision ou le rappel avec une forte probabilité est par conséquent de rechercher les valeurs de θ' les plus proches possibles de cette barrière.

Le théorème ci-dessous permet d'établir les valeurs des supports statistiques que nous proposons en calculant la valeur de ε :

Théorème 1 $\forall X, \forall \mathcal{D}, \forall m > 0, \forall 0 \leq \theta \leq 1, \forall 0 < \delta \leq 1$, *on choisit* ε *tel que :*

$$\varepsilon \quad \geq \quad \sqrt{\frac{1}{2m} \ln \frac{|S^*|}{\delta}} \ .$$

Si on fixe $\theta' = \theta + \varepsilon$*, alors* $\mathtt{P} = 1$ *avec une probabilité au moins de* $1 - \delta$*. Si on fixe* $\theta' = \theta - \varepsilon$*, alors* $\mathtt{R} = 1$ *avec une probabilité au moins de* $1 - \delta$*.* δ *est le risque statistique lié aux data streams. Les valeurs* $\theta + \varepsilon$*,* $\theta - \varepsilon$ *sont les supports statistiques au sens de la définition 1.*

Les supports obtenus ($\theta' = \theta \pm \varepsilon$) sont statistiquement presque optimaux. Faute de place, nous renvoyons à l'article (Nock et al., 2005) [1] pour la démonstration complète. Celle-ci repose sur l'utilisation d'inégalités de concentration de variables aléatoires, qui, dans ce cas préçis, permettent d'obtenir des résultats statistiquement presque optimaux. Par optimalité, nous entendons que toute technique d'estimation obtenant de meilleures bornes est condamnée à se tromper (le critère à maximiser n'est plus ègal à un) quelque soit son temps de calcul.

3 Bordures statistiques pour la mise à jour incrémentale

Dans cette section nous introduisons deux bordures statistiques (supérieure et inférieure) qui vont être pertinentes dans le choix des motifs fréquents à conserver lors de la mise à jour incrémentale. L'objectif avec la bordure statistique inférieure est de maximiser la précision $\mathtt{P}$, tandis qu'avec la bordure statistique supérieure il s'agit de maximiser le rappel $\mathtt{R}$. Nous adoptons la notation probabiliste définie par (McAllester, 1999).[2]

A partir du théorème 1, nous définissons l'ensemble suivant : pour un risque statistique δ fixé par l'utilisteur, en choisissant $\theta' = \theta + \varepsilon$, on construit l'ensemble $S^*_{\theta+\varepsilon}$ tel que $\forall^\delta, S^*_{\theta+\varepsilon} \subseteq X^*_\theta$ avec $\forall^\delta, \mathtt{P} = 1$. Ainsi, il n'y a plus la première source d'erreurs avec une forte probabilité. Tous les motifs de $S^*_{\theta+\varepsilon}$ sont des motifs, vrais *θ-fréquents* de X^*_θ, mais on ne les a pas tous (voir figure 3). $S^*_{\theta+\varepsilon}$ est le plus grand ensemble possible qui contient uniquement des motifs vrais *θ-fréquents* de X^*_θ après un temps d'observation du stream. Nous définissons cet ensemble comme étant la bordure statistique inférieure de X^*_θ à partir de l'échantillon des motifs S^* de X. De façon symétrique, à partir du théorème 1, nous définissons l'ensemble suivant : pour un

[1] www.pa-laur.com/LNSP_CIKM05.pdf

[2] Cette notation concise, s'écrivant $\forall^\delta P$, précise que : le prédicat P est vrai à une fraction près $\leq \delta$ pour l'ensemble S obtenu à partir de la distribution $\mathcal{D}$. De façon équivalente, cela signifie que le prédicat P est vrai avec une probabilité $\geq 1 - \delta$ pour l'ensemble S obtenu à partir de la distribution D.

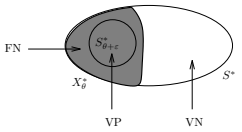

FIG. 3 – *Bordure statistique inférieure.*

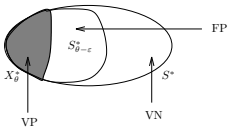

FIG. 4 – *Bordure statistique supérieure.*

risque statistique δ fixé par l'utilisteur, en choisissant $\theta' = \theta - \varepsilon$, on construit l'ensemble $S^*_{\theta-\varepsilon}$ tel que $\forall^\delta, X^*_\theta \subseteq S^*_{\theta-\varepsilon}$ avec $\forall^\delta, \mathtt{R} = 1$. Ainsi, il n'y a plus la deuxième source d'erreurs avec une forte probabilité, c'est à dire que $S^*_{\theta-\varepsilon}$ contient tous les motifs, vrais θ-*fréquents* de X^*_θ, mais il en contient d'autres (figure 4). $S^*_{\theta-\varepsilon}$ est le plus petit ensemble possible qui contient tous les motifs vrais θ-*fréquents* de X^*_θ après un temps d'observation du stream. Nous définissons cet ensemble comme étant la bordure statistique supérieure de X^*_θ à partir de l'échantillon des motifs S^* du stream X.

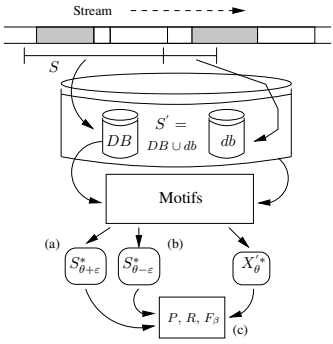

FIG. 5 – *Bordures statistiques pour la mise à jour incrémentale.*

La figure 5 illustre l'utilisation de bordures dans un processus de mise à jour incrémentale. Nous recherchons les motifs θ-*fréquents* à partir des bordures statistiques (notées (a) et (b) sur la figure 5). Ces bordures sont construites pour le support θ choisi à partir de la valeur de ε, calculée pour S (DB). Ces bordures, nous permettent d'approcher au mieux l'ensemble de tous les motifs vrais θ-*fréquents* pour S', qui représente la base DB mise à jour par l'ajout de db, $S' = S \cup db$. Dans la section suivante lors de nos expérimentations, nous comparerons les motifs θ-*fréquents* obtenus grâce aux bordures statistiques par rapport à l'ensemble X'^*_θ ((c) sur la figure 5). Cet ensemble représente les vrais motifs θ-*fréquents* après mise à jour incrémentale ($DB \cup db$).

4 Expérimentation

Comme nous l'avons précisé dans la section précédente notre objectif lors des expérimentations ne consiste pas à évaluer les temps de réponse associés à nos bordures mais plutôt

Database	θ	taille DB	taille db
Dragons	[0.07, 0.2] / 0.03	[0.2,0.6] / 0.1	[0.1, 0.5] /0.1
BuAG	[0.08, 0.2] / 0.03	[0.2,0.6] / 0.1	[0.1, 0.5] /0.1

FIG. 6 – *Valeurs des paramètres de la forme $[a, b]/c$, où a est la valeur de départ, c est l'incrément, et b est la valeur finale.*

d'évaluer leur qualité. Pour cela, nous utilisons les mesures de précision P et de rappel R définies précédemment dans les équations (1) et (2). Du fait de l'indépendance de notre méthode vis à vis des motifs recherchés, nous avons utilisé un algorithme traditionnel de recherche de motifs séquentiels fréquents SPADE (Zaki, 2001).

Lors de cette évaluation, nous avons utilisé deux jeux de données issus de serveurs Web. Le premier appelé "Dragons", est obtenu sur le site internet[3]. Ces données représentent la navigation d'utilisateurs sur ce site. La taille du fichier de log représente environ 2,54Go (132k transactions). La deuxième base de données, appelée "BuAG", est obtenue à partir des 3,48Go (54k transactions) de Web log du serveur web de la bibliothèque de l'Université[4].

Pour analyser la qualité de nos bordures stastiques, nous évaluons différentes situations en faisant varier un ensemble de paramètres (une exception est faite pour δ, qui est fixé à .05). Les variations de ces paramètres sont décrites dans la figure 6. Le premier paramètre définit les différentes valeurs de support sur lesquelles nous allons réaliser l'expérimentation ("θ"). Le second paramètre, définit la taille de DB par rapport à la taille du stream simulé ("taille DB"). Le dernier paramètre, quant à lui, définit la taille de db par rapport à celle de DB. Dans notre cas db représentera au plus 50% de DB ("taille db"), ce paramètre permet de contrôler la taille de l'incrément par rapport à la partie stockée initialement. Pour gérer et organiser ces expériences un générateur est chargé de coordonner tous les tests à effectuer.

Au lieu d'utiliser un vrai data stream, qui aurait pu limiter la qualité de l'évaluation de nos bordures statistiques, nous avons choisi d'en simuler un à l'aide de la connaissance de son domaine X. Plus précisément, pour simuler le stream nous échantillonnons chaque base de données en fragments DB (S). Par exemple, nous pouvons considérer que les données arrivent successivement depuis la base de données "Dragons" et que nous ne pouvons en stocker que 20%. Pour cela nous prenons 20% des transactions contenues dans cette base de données et nous les conservons dans DB. Il ne reste alors plus qu'à prendre un incrément, db par exemple de taille 10%, de cette base. Nous construisons alors les bordures statistiques (bordure inférieure et bordure supérieure) définies dans la section 3 pour DB.

Les figures 7 et 8 montrent les résultats d'expériences obtenues, avec $\delta = 0.05$, respectivement sur les bases Dragons et BuAG. Pour évaluer la qualité de ces bordures pour la mise à jour, nous représentons leurs comportements pour P et R par rapport à $S^{'}$ ($S^{'} = DB \cup db$). Ainsi, les courbes $P_{\theta+\varepsilon}$ et $P_{\theta-\varepsilon}$, représentent la précision respectivement pour la bordure statistique inférieure et pour la bordure statistique supérieure. De même les courbes $R_{\theta\pm\varepsilon}$ représentent le rappel. Les courbes P_θ et R_θ correspondent au choix trivial $\theta' = \theta$. Nous constatons que ces courbes ont un comportement assez similaire :

– la précision P vaut ou approche 1 pour la plupart des bases stockées quand $\theta^{'} = \theta + \varepsilon$,

[3] www.elevezundragon.com.
[4] www.univ-ag.fr/buag/.

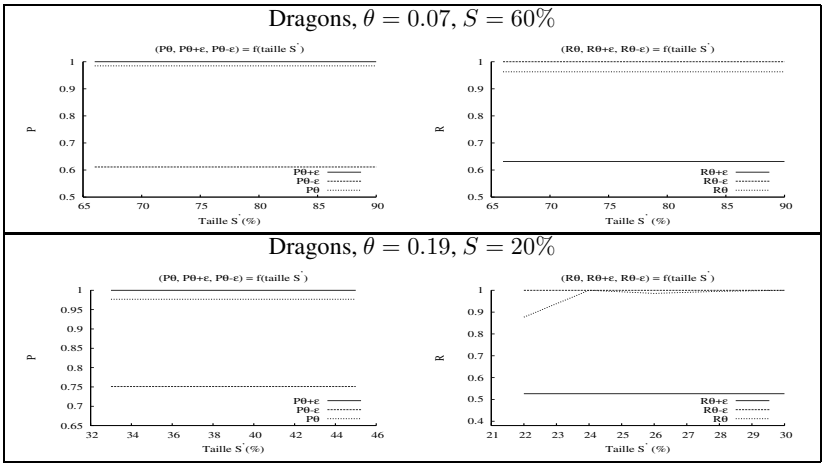

FIG. 7 – *représente les courbes* P *(à gauche) et* R *(à droite) sur la base Dragons pour trois valeurs de* θ' *:* $\theta - \varepsilon, \theta$ *et* $\theta + \varepsilon$*, deux valeurs de* θ *et deux tailles de* S *(*$\%$ *par rapport à* $|X|$*).*

– le rappel R vaut ou approche 1 pour la plupart des bases stockées quand $\theta^{'} = \theta - \varepsilon$.

Ces observations sont en accords avec les résultats théoriques de la section 2. Nous observons également un autre phénomène : le rappel R associé à $\theta^{'} = \theta + \varepsilon$ n'est pas très éloigné de celui associé à $\theta^{'} = \theta$. Il en est de même pour la précision P associée à $\theta^{'} = \theta - \varepsilon$. Ce qui montre que la maximisation de la précision P ou du rappel R est obtenue avec un coût réduit au niveau de la dégradation de l'autre paramètre. Nous notons aussi que les courbes de précision P ont de meilleures performances que celles du rappel R notamment sur la figure 8. Ceci n'est pas très surprenant ($c.f.$ section 2), car la fourchette de valeur pour la précision P est beaucoup plus restreinte que pour le rappel R.

La variation de la taille de S (DB) possède également une importance et vient conforter un résultat théorique auquel on pouvait s'attendre. En effet si on s'intéresse au rappel R, on constate, que ce soit sur les figures 7 ou 8, que les valeurs observées pour cette quantitée augmentent avec la taille de S. Cela traduit le fait que plus nous possédons de données observées plus la qualité de la prédiction augmente.

Sur ces bases de données un autre phénomène semble apparaître. Premièrement, à cause des faibles valeurs de θ, certains tests n'ont pas pu être réalisés car la valeur de $\theta - \varepsilon$ était inférieure à 0. De plus, les différences observées entre les courbes semblent être liées aux tailles des bases de données utilisées. La base de données $BuAG$ est plus petite que celle de $Dragons$ d'un facteur 2.4. Nous pensons que ceci explique la différence entre les courbes : il s'agit de phénomènes liés à des bases de données de petites tailles et qui ne devraient pas être présents dans des bases de données plus conséquentes ou dans de vrais data streams.

Sur ces courbes (figures 7 et 8), le choix de $\theta' = \theta$ donne de meilleurs résultats en ce qui concerne la moyenne de P et R que le choix des deux valeurs de θ', sachant que ni P, ni R ne sont très proches de 1 avec une forte probabilité dans ce cas. Avec notre approche, on peut efficacement optimiser le processus de mise à jour incrémentale. Dans le cas où on aurait un

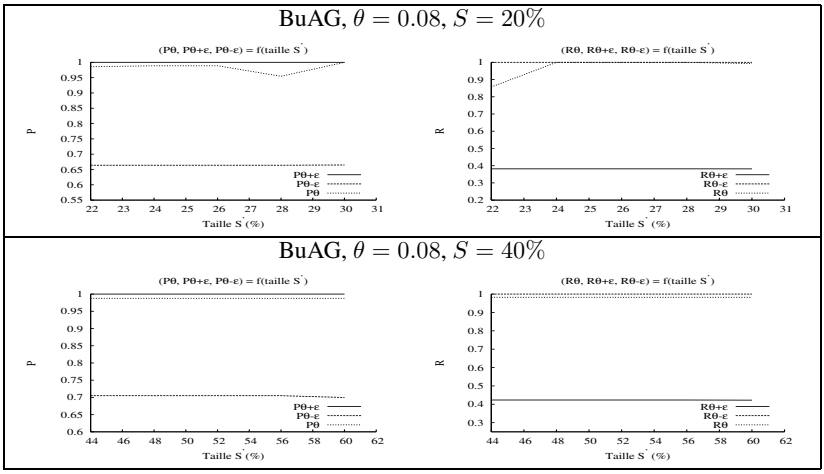

FIG. 8 – *représente les courbes* P *(à gauche) et* R *(à droite) sur la base BuAG pour trois valeurs de* θ' *:* $\theta - \varepsilon, \theta$ *et* $\theta + \varepsilon$*, deux valeurs de* θ *et deux tailles de* S *(*$\%$ *par rapport à* $|X|$*).*

espace suffisant de stockage, on choisirait $\theta' = \theta - \varepsilon$, la bordure supérieure, pour laquelle on possède des garanties importantes sur la présence des futurs fréquents en son sein ($\mathtt{R} = 1$). Dans ce cas, le nombre de calculs supplémentaires sera faible lors de la mise à jour. Par contre dans le cas où on devrait se limiter pour des raisons de taille à conserver une bordure disons moins volumineuse, $\theta' = \theta + \varepsilon$, la bordure inférieure, sera alors utilisée car elle contient les informations les plus pertinentes ($\mathtt{P} = 1$). On peut ainsi voir les bordures statistiques comme des outils de mise à jour incrémentale indiquant les limites au delà desquelles : soit il est inutile de stocker plus d'informations dans le cas de $\theta' = \theta - \varepsilon$ (valeur limite inférieure pour $\theta^{'}$) ; soit la perte d'information serait trop importante $\theta' = \theta + \varepsilon$ (valeur limite supérieure pour $\theta^{'}$).

5 Travaux connexes

Depuis 1996, de nombreux travaux de recherche se sont focalisés sur la maintenance des ensembles de motifs fréquents obtenus dans les bases de données statiques. Dans ce paragraphe nous regardons leur adéquation par rapport à notre problématique. Partition et FUP (Fast UPdate) (Cheung et al., 1996) sont deux algorithmes où un partitionnement de la base de données est effectué pour rechercher dans un premier temps les itemsets fréquents locaux relatifs à chaque partition puis dans un second temps en déduire les itemsets fréquents globaux par validation croisée. Cela repose sur l'hypothèse que les itemsets fréquents de la base doivent l'être dans au moins l'une des partitions. (Parthasarathy et al., 1999) ont développé un algorithme de mise à jour incrémentale ISM (Incremental Sequence Mining) en maintenant un treillis, de motifs de la base, construit à partir de tous les motifs fréquents et de tous les motifs de la bordure négative. (Zheng et al., 2002) ont développé un algorithme IUS(Incrementally Updating Sequence) en utilisant une valeur support pour limiter la taille de l'espace des candi-

dats de la bordure négative. Ces différentes approches se heurtent aux inconvénients inhérents à l'utilisation de la bordure négative :

– l'espace des motifs candidats à maintenir est très important ;
– il est nécessaire de considérer les relations structurelles qui existent entre les motifs notamment dans le cas des motifs qui auraient une faible valeur de support.

Les différents algorithmes utilisant la bordure négative sont très coûteux en temps et sont consommateurs d'espace mémoire.

(Masseglia et al., 2003) ont développé un algorithme de mise à jour incrémentale ISE utilisant une approche par génération et test de candidats. Les inconvénients sont que :

– l'espace des candidats peut être très grand, ce qui rend la phase de test très lente ;
– l'algorithme requiert plusieurs passages sur toute la base. Cela est très coûteux en temps particulièrement pour les motifs séquentiels à séquences longues.

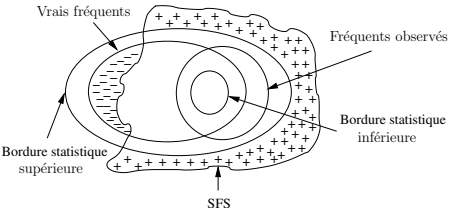

FIG. 9 – *Comparaison.*

L'algorithme IncSpan (Incremental Mining of Sequential Patterns in large database) développé par (Cheng et al., 2004) repose sur une approche statistique où l'on construit un ensemble de motifs semi-fréquents (SFS) en diminuant la valeur du support à partir d'un ratio. L'idée est de dire qu'à partir d'un ensemble de motifs "presque fréquents" (SFS), plusieurs des motifs fréquents de la base mise à jour proviendraient de SFS ou alors seraient déjà observés fréquents dans la base connue précédant la mise à jour. Autrement dit, SFS constituerait une zone frontière entre les motifs fréquents et non fréquents. Sur la figure 9, nous représentons un exemple d'ensemble SFS. Cette approche présente des insuffisances majeures d'un point de vue statistique tant pour l'estimation de l'incertitude intrinsèque liée aux data streams que pour la construction de l'ensemble des motifs candidats permettant la mise à jour incrémentale. En effet, pour diminuer la valeur du support, le choix du ratio est simplement heuristique sans justification théorique. Ainsi, cette approche n'offre aucune certitude ni indication sur l'erreur commise lors de la construction de l'ensemble des motifs semi-fréquents quant aux motifs vrais θ*-fréquents* du stream (zone identifiée avec les symboles - sur la figure 9). De plus, nous n'avons aucune garantie sur la minimalité de cet ensemble (zone identifiée avec les symboles + sur la figure 9), ce qui est fortement pénalisant pour sa réutilisation dans l'objectif d'optimisation de la méthode de mise à jour incrémentale.

6 Conclusion

Dans cet article, nous abordons la problématique de la mise à jour incrémentale pour les motifs fréquents dans le cas des grandes bases de données. Dans le contexte des data streams,

il est plus pertinent, de construire et de maintenir des ensembles de motifs vrais *θ-fréquents* et non simplement observés comme tels. Plusieurs travaux, (Kearns et Mansour, 1998; Nock et Nielsen, 2004; Vapnik, 1998), ont montré l'intérêt de l'approche statistique notamment dans la détermination de règles de prévision pour l'optimisation d'algorithmes. Notre contribution majeure porte précisément sur ce point par l'introduction de supports statistiques en complément des supports classiques permettant d'obtenir des bordures statistiques qui constituent les ensembles statistiquement presque optimaux (à une constante près) à considérer dans le cadre de la mise à jour incrémentale. Les expérimentations présentées montrent la robustesse de l'approche, dans le cas des motifs séquentiels, au regard de la taille des bases stockées et des différentes valeurs supports testées. Ces résultats encourageants sont des points positifs quant à l'applicabilité et le passage à l'échelle de la méthode. Plusieurs extensions de ces travaux sont possibles dans bien des domaines de recherche en fouille de données. On citera notamment les travaux qui portent sur les structures de données où l'on cherche à maintenir des ensembles d'items qui sont observés fréquents avec un rappel maximum, (Jin et al., 2003). Nos préoccupations actuelles concernent la question suivante : est-il possible de construire un ensemble intermédiaire qui conserverait au mieux les propriétés de chacune de ces bordures ? Celui-ci représenterait un compromis entre une bordure statistique trop imposante à stocker et une autre pour laquelle le nombre d'accès à la base de données est trop important.

Références

Agrawal, R., T. Imielinski, et A.-N. Swami (1993). Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD ICMD'93*, pp. 207–216.

Ayres, J., J. Flannick, J. Gehrke, et T. Yiu (2002). Sequential pattern mining using bitmap representation. In *Proc. of KDD'02*.

Cheng, X., X. Yan, et J. Han (2004). Incspan : Incremental mining of sequential patterns in large database. In *Proc. of KDD'04*.

Cheung, D., J. Han, V. Ng, et C. Wong (1996). Maintenance of discovered association rules in large databases : an incremental updating technique. In *Proc. of ICDE'96*, pp. 106–114.

Demaine, E., A. Lopez-Ortizand, et J.-I. Munro (2002). Frequency estimation of internet packet streams with limited space. In *Proc. of the 10^{th} European Symposium on Algorithms*.

Devroye, L., L. Györfi, et G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*.

Fan, W., Y.-A. Huang, H. Wang, et P.-S. Yu (2004). Active mining of data streams. In *Proc. of the 4^{th} SIAM International Conference on Data Mining*, pp. 457–461.

Golab, L. et M. T. Ozsu (2003). Issues in data stream management. *ACM SIGMOD Records 2*.

Han, J., J. Pei, B. Mortazavi-asl, Q. Chen, U. Dayal, et M. Hsu (2000). Freespan : Frequent pattern-projected sequential pattern mining. In *Proc. of KDD'00*.

Jin, C., W. Qian, C. Sha, J.-X. Yu, et A. Zhou (2003). Dynamically maintaining frequent items over a data stream. In *Proc. of CIKM'03*, pp. 287–294. ACM Press.

Karp, R.-M., S. Shenker, et C.-H. Papadimitriou (2003). A simple algorithm for finding elements in streams and bags. *ACM Trans. on Database Systems 28*, 51–55.

Kearns, M. J. et Y. Mansour (1998). A Fast, Bottom-up Decision Tree Pruning algorithm with Near-Optimal generalization. In *Proc. of ICML'98*, pp. 269–277.

Li, H.-F., S.-Y. Lee, et M.-K. Shan (2004). An efficient algorithm for mining frequent itemsets over the entire history of data streams. In *Proc. of the 1^{st} Int. Workshop on Knowledge Discovery in Data Streams*.

Manku, G. et R. Motwani (2002). Approximate frequency counts over data streams. In *Proc. of the 28 th International Conference on Very Large Databases*, pp. 346–357.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowlede discovery. *Data Mining and Knowledge Discovery 1*, 241–258.

Masseglia, F., P. Poncelet, et M. Teisseire (2003). Incremental mining of sequential patterns in large databases. *Data and Knowledge Engineering 46*.

McAllester, D. (1999). Some PAC-Bayesian theorems. *Machine Learning 37*, 355–363.

Nock, R., P. Laur, P. Poncelet, et J. Symphor (2005). On the estimation of frequent itemsets for data streams : Theory and experiments. In *Proc. of CIKM'05*.

Nock, R. et F. Nielsen (2004). Statistical Region Merging. *IEEE Trans. on Pattern Analysis and Machine Intelligence 26*, 1452–1458.

Parthasarathy, S., M. Zaki, M. Orihara, et S. Dwarkadas (1999). Incremental and interactive sequence mining. In *Proc. of CIKM'99*.

Pei, J., J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen, et U. Dayal (2001). Prefixspan : Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. of ICDE'01*.

Srikant, R. A. R. (1995). Mining sequential patterns. In *Proc. of ICDE'95 Conference*.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley.

Wang, H., W. Fan, P.-S. Yu, et J. Han (2003). Mining concept-drifting data streams with ensemble classifiers. In *Proc. of KDD'03*, pp. 226–235.

Zaki, M. (2001). SPADE : An efficient algorithm for mining frequent sequences. *Machine Learning Journal 42*.

Zheng, Q., K. Xu, S. Ma, et W. lv (2002). The algorithms of updating sequential patterns. In *Proc. of ICDM'02*.

Summary

Recently the knowledge extraction community takes a closer look to new models where data arrive in timely manner like a fast and continous flow, $i.e.$ data streams. One of the most important singularity inside streams relies on that only a part of it is available. After some following updates, it's necessary, to cope with uncertainty as only a part of it is available. In this paper, we introduce a new statiscal approach which biases the initial support for sequential patterns mining. This approach holds the advantage to maximize one of the parameters (precision or recall) chosen by the user while limiting the degradation of the other criterion. Thus, we define the statistical borders which are the relevant sets of frequent patterns in incremental mining of streams. Experiments performed on sequential patterns demonstrate the interest of this approach and the potential of such techniques.