Equilibrer l'analyse des motifs fréquents

Arnaud Giacometti, Patrick Marcel, Arnaud Soulet

Université François Rabelais Tours, LI
3 place Jean Jaurès
F-41029 Blois France
prenom.nom@univ-tours.fr

Résumé. Cet article propose une méthode originale d'évaluation de la qualité des motifs en anticipant la manière qui sera utilisée pour les analyser. Nous commençons par introduire le modèle de l'analyse aléatoire d'un ensemble de motifs selon une mesure d'intérêt. Avec ce modèle, nous constatons que l'étude des motifs fréquents avec le support conduit à une analyse déséquilibrée du jeu de données. Afin que chaque transaction reçoive la même attention, nous définissons le support équilibré qui corrige le support classique en pondérant les transactions. Nous proposons alors un algorithme qui calcule ces poids et nous validons expérimentalement son efficacité.

1 Introduction

La découverte de motifs introduite par Agrawal et Srikant (1994) consiste à extraire des informations pertinentes décrivant une partie des données. Depuis une quinzaine d'années, les algorithmes ont gagné en performance et arrivent désormais à extraire rapidement les motifs depuis des données volumineuses. Cependant, évaluer et garantir la qualité des motifs extraits demeure une problématique très ouverte. On distingue dans la littérature deux approches : celles guidées par les données (évaluant l'intérêt des motifs sur les données à analyser) et celles guidées par l'utilisateur (bénéficiant d'informations issues de l'utilisateur). Dans cet article, nous souhaitons adopter une nouvelle approche, dite *guidée par l'analyse*. L'évaluation en amont de l'intérêt des motifs s'appuie alors sur la manière dont les motifs seront analysés.

$\mathcal{D}$

Tid	Items
t_1	A B
t_2	A B
t_3	A B
t_4	C
t_5	C
t_6	C

+

P

Pid	Itemset	Support	Proportion d'analyse
p_1	A	0.5	0.5/2
p_2	B	0.5	0.5/2
p_3	AB	0.5	0.5/2
p_4	C	0.5	0.5/2

$\rightarrow$

Répartition de l'analyse

Tid	Prop. d'analyse
t_1	0.75
t_2	0.75
t_3	0.75
t_4	0.25
t_5	0.25
t_6	0.25

TAB. 1 – *Une analyse* déséquilibrée *du jeu de données $\mathcal{D}$ avec les motifs fréquents P*

Illustrons notre démarche sur un exemple jouet. Le tableau 1 présente un jeu de données $\mathcal{D}$ contenant 6 transactions composées des items A, B et C ainsi que les 4 itemsets présents

dans au moins 50% des transactions de $\mathcal{D}$. Supposons qu'un analyste s'appuie sur les motifs P pour étudier le jeu de données $\mathcal{D}$. S'il consacre autant de temps à chacun des motifs, 75% de son analyse portera sur des motifs décrivant la première moitié de $\mathcal{D}$ tandis que seulement 25% sera dédiée au dernier motif de P, le seul à couvrir la seconde moitié de $\mathcal{D}$. Le poids final de t_1, t_2 ou t_3 dans l'analyse est donc supérieur à celui de chacune des transactions t_4, t_5 ou t_6. Nous dirons alors que l'analyse du jeu de données est déséquilibrée. Dans la suite, nous cherchons à rééquilibrer l'analyse en renforçant les transactions qui sont les moins décrites. A notre connaissance, ce problème n'est pas abordé dans la littérature même si les représentations condensées (Pasquier et al., 1999) en éliminant les redondances réduisent le déséquilibre de l'analyse. De manière plus significative, les modèles globaux fondés sur des motifs (Fürnkranz et Knobbe, 2010) favorisent des analyses équilibrées. Malheureusement ces modèles éliminent de nombreux motifs pertinents.

Cet article vise à vérifier si intégrer la méthode d'analyse des motifs lors de l'évaluation de leur intérêt améliore la pertinence des motifs découverts. Plus précisément, notre première contribution est la proposition du *modèle de l'analyse aléatoire* qui simule des sessions d'analyse d'un ensemble de motifs en fonction d'une mesure d'intérêt (cf. la section 3). Nous définissons alors la notion d'analyse équilibrée où chaque transaction est étudiée avec la même acuité. Notre seconde contribution présentée à la section 4 est l'introduction d'une nouvelle mesure d'intérêt, appelée *support équilibré* qui selon notre modèle corrige le support pour induire une analyse équilibrée des motifs fréquents. Nous proposons alors un algorithme, nommé SUPPORTBALANCE, afin de calculer cette nouvelle mesure. Enfin, une étude expérimentale à la section 5 valide l'efficacité de SUPPORTBALANCE.

2 Notations

Cet article s'appuie sur le cadre de Mannila et Toivonen (1997). Un *langage* $\mathcal{L}$ est un ensemble de motifs. Par exemple, dans le tableau 1, le jeu de données $\mathcal{D}$ est un multi-ensemble du langage d'itemsets. Une *relation de spécialisation* $\preceq$ est un ordre partiel sur $\mathcal{L}$. Si $\preceq$ est une relation de spécialisation sur $\mathcal{L}$, $l \preceq l'$ signifie que l est plus général que l' et l' est plus spécifique que l. Par exemple, A est plus général que AB suivant la spécialisation $\subseteq$.

Comme il est parfois nécessaire de mettre en relation des langages distincts (e.g., pour relier des motifs aux données), on utilise la notion de couverture. Une *relation de couverture* est une relation binaire $\triangleleft \subseteq \mathcal{L}_1 \times \mathcal{L}_2$ (où $\mathcal{L}_1$ et $\mathcal{L}_2$ sont deux langages) ssi quand $l_1 \triangleleft l_2$, on a $l'_1 \triangleleft l_2$ (resp. $l_1 \triangleleft l'_2$) pour n'importe quel motif $l'_1 \preceq l_1$ (resp. $l_2 \preceq l'_2$). La relation $l_1 \triangleleft l_2$ signifie que l_1 couvre l_2 et l_2 est couvert par l_1. Dans le tableau 1, la relation d'inclusion est par exemple utilisée pour déterminer les transactions de $\mathcal{D}$ couvertes par un motif de P. Etant donné deux ensembles de motifs $L \subseteq \mathcal{L}$, $L' \subseteq \mathcal{L}'$ et une relation de couverture $\triangleleft \subseteq \mathcal{L} \times \mathcal{L}'$, les *motifs couverts* de L' par $l \in L$ est l'ensemble des motifs de L' couverts par le motif l : $L'_{\triangleright l} = \{l' \in L' | l \triangleleft l'\}$. De manière duale, les *motifs couvrants* de L pour $l' \in L'$ est l'ensemble des motifs de L couvrant le motif l' : $L_{\triangleleft l'} = \{l \in L | l \triangleleft l'\}$. Avec le tableau 1, on obtient $\mathcal{D}_{\supseteq AB} = \{t_1, t_2, t_3\}$ et $P_{\subseteq t_1} = \{A, B, AB\}$.

Afin d'évaluer la pertinence d'un motif, les processus d'extraction de motifs exploitent des mesures d'intérêt. Typiquement, le support d'un motif φ dans le jeu de données $\mathcal{D}$ est le nombre de transactions couvertes par φ (Agrawal et Srikant, 1994) : $Supp(\varphi, \mathcal{D}) = |\mathcal{D}_{\triangleright\varphi}|/|\mathcal{D}|$. Un motif est alors dit fréquent lorsque son support excède un seuil minimal spécifié par l'utilisateur.

Par exemple, avec un seuil minimal de 0.5, le motif AB est fréquent car $\mathit{Supp}(AB, \mathcal{D}) = |\{t_1, t_2, t_3\}|/6 \geq 0.5$. Dans la suite, toute fonction $f : \mathcal{L} \to \mathbb{R}$ est étendue en considérant que $f(P) = \sum_{\varphi \in P} f(\varphi)$ pour tout P multi-ensemble de $\mathcal{L}$. De cette manière, on a $\mathit{Supp}(P, \mathcal{D}) = \sum_{\varphi \in P} \mathit{Supp}(\varphi, \mathcal{D})$ pour tout ensemble de motifs P.

3 Modèle de l'analyse aléatoire d'un ensenble de motifs

Définition du modèle de l'analyse aléatoire Notre travail repose sur l'idée d'intégrer la méthode d'analyse des motifs dès l'évaluation des motifs. Pour cela, nous modélisons l'analyse d'un ensemble de motifs à l'instar du modèle du surfeur aléatoire (Brin et Page, 1998).

Le *modèle de l'analyse aléatoire* génère des sessions en tirant aléatoirement des motifs. Plus précisément, l'« analyste modélisé » tire au hasard un motif en favorisant ceux de plus forte mesure. Il étudie alors ce motif et les transactions couvertes par ce dernier pendant un laps constant. Après chaque analyse de motif, la session peut soit s'interrompre (si l'analyste est satisfait), soit se poursuivre (si l'analyste est insatisfait).

Définition 1 (Modèle de l'analyse aléatoire) *Soient un jeu de données $\mathcal{D}$, un ensemble de motifs $P \subseteq \mathcal{L}$ et une mesure d'intérêt $m : \mathcal{L} \to [0, 1]$. Le modèle de l'analyse aléatoire avec une probabilité d'arrêt $\alpha \in]0, 1[$ et une durée $\delta > 0$, noté $\mathcal{A}_{\alpha,\delta}$, génère une session avec le processus suivant :*

1. *Tirer un motif φ de P suivant la distribution $p(\gamma) = m(\gamma)/m(P)$ (où $\gamma \in P$).*
2. *Etudier φ et les transactions de $\mathcal{D}$ couvertes par φ pendant une durée δ.*
3. *Stopper la session avec une probabilité α ou alors, poursuivre à l'étape 1.*

Nous utilisons à présent le modèle de l'analyse aléatoire pour dériver des informations en moyenne sur les analyses. Par exemple, la probabilité qu'une session soit de longueur $k > 0$ est $\alpha \times (1 - \alpha)^{(k-1)}$. On en déduit que la longueur moyenne des sessions générées par $\mathcal{A}_{\alpha,\delta}$ est $\sum_{k>0} k \times \alpha \times (1 - \alpha)^{(k-1)}$ qui est égal à $1/\alpha$. Comme on consacre δ temps à l'étude de chaque motif tiré, la durée moyenne des sessions correspond à δ/α. Etant donné que les motifs sont tirés suivant la distribution p, la durée moyenne d'analyse d'un motif φ, notée $\Delta(\varphi, \mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, m))$, est proportionnelle à sa probabilité de tirage : $\Delta(\varphi, \mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, m)) = p(\varphi) \times \delta/\alpha = m(\varphi)/m(P) \times \delta/\alpha$. Enfin, une transaction est étudiée à chaque fois qu'un motif qui la couvre est étudié. On obtient alors pour tout $t \in \mathcal{D}$:

$$\Delta(t, \mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, m)) = \frac{m(P_{\lhd t})}{m(P)} \times \delta/\alpha \tag{1}$$

Equilibre de l'analyse Une analyse est pertinente si elle rend compte de l'ensemble des transactions avec la même acuité. Plus formellement, une analyse est équilibrée lorsque la durée d'analyse de chaque transaction t est égale à la durée moyenne : $\Delta(t, \mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, m)) = \sum_{t' \in \mathcal{D}} \Delta(t', \mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, m))/|\mathcal{D}|$. En injectant l'équation 1 et en éliminant le facteur $1/m(P) \times \delta/\alpha$ à gauche et à droite, on obtient pour chaque transaction $t \in \mathcal{D}$:

$$m(P_{\lhd t}) = \frac{1}{|\mathcal{D}|} \times \sum_{t' \in \mathcal{D}} m(P_{\lhd t'}) \tag{2}$$

Cette équation est vérifiée quand $m(P_{\lhd t})$ est constant pour toute transaction t. Comme le montre l'exemple jouet du tableau 1, l'analyse des motifs fréquents d'un jeu de données avec la mesure de support est en général déséquilibrée. En effet, les motifs fréquents ont tendance à concentrer l'analyse sur les transactions les plus communes. Ainsi, des phénomènes plus rares et recoupant des transactions marginales risquent d'être entièrement occultés.

4 Equilibrer l'analyse des motifs fréquents

$\mathcal{D}_b$

Tid	Items
t_1	A B
t_2	C
t_3	C
t_4	C

+

P

Pid	Itemset	Support	Proportion d'analyse
p_1	A	0.25	0.25/1.5
p_2	B	0.25	0.25/1.5
p_3	AB	0.25	0.25/1.5
p_4	C	0.75	0.75/1.5

$\rightarrow$

Répartition de l'analyse

Tid	Prop. d'analyse
t_1	0.5
t_2	0.5
t_3	0.5
t_4	0.5

TAB. 2 – *Une analyse* équilibrée *du jeu de données $\mathcal{D}_b$ avec les motifs fréquents P*

Support pondéré Nous choisissons de modifier le support afin d'équilibrer $\mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, \mathit{Supp})$ sans perturber ni le jeu de données initial $\mathcal{D}$, ni l'ensemble de motifs étudié P. L'analyse des motifs fréquents P du jeu de données $\mathcal{D}_b$ présentée par le tableau 2 est équilibrée. Intuitivement, cet équilibre découle d'un renforcement des transactions de $\mathcal{D}$ qui sont les moins décrites (i.e., C, cf. tableau 1). Afin de simuler l'évolution de $\mathcal{D}$ à $\mathcal{D}_b$ (sans construire $\mathcal{D}_b$), seule une pondération des transactions peut être introduite dans le calcul du support (dit alors *pondéré*) :

Définition 2 (Support pondéré) *Le support pondéré par $w : \mathcal{D} \to \mathbb{R}_+^*$ d'un motif φ dans le jeu de données $\mathcal{D}$ est défini par $\mathit{Supp}_w(\varphi, \mathcal{D}) = w(\mathcal{D}_{\rhd\varphi})/w(\mathcal{D})$.*

Par exemple, avec la pondération b où $t_1, t_2, t_3 \mapsto 1/12$ et $t_4, t_5, t_6 \mapsto 1/4$, on obtient $\mathit{Supp}_b(AB, \mathcal{D}) = 1/12 + 1/12 + 1/12 = 0.25$ et $\mathit{Supp}_b(C, \mathcal{D}) = 1/4 + 1/4 + 1/4 = 0.75$ (car $b(\mathcal{D}) = 1$). On constate que le support pondéré par la fonction $t \mapsto 1/|\mathcal{D}|$ correspond exactement au support traditionnel. Plus généralement, le support pondéré dans un jeu de données $\mathcal{D}$ est équivalent au support classique dans un jeu de données où la présence des transactions est pondérée par leur poids. Par exemple, on verifie bien que $\mathit{Supp}_b(AB, \mathcal{D}) = \mathit{Supp}(AB, \mathcal{D}_b)$ et $\mathit{Supp}_b(C, \mathcal{D}) = \mathit{Supp}(C, \mathcal{D}_b)$ avec les exemples des tableaux 1 et 2.

Support équilibré En injectant le support pondéré par w des motifs P dans $\mathcal{D}$ dans l'équation 2, une analyse est équilibrée ssi pour tout $t \in \mathcal{D}$:

$$\mathit{Supp}_w(P_{\lhd t}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \times \sum_{t' \in \mathcal{D}} \mathit{Supp}_w(P_{\lhd t'}, \mathcal{D}) \tag{3}$$

L'algorithme ci-dessous SUPPORTBALANCE retourne une pondération w pour équilibrer au mieux l'analyse $\mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, \mathit{Supp}_w)$ car l'équation 3 n'admet pas toujours de solution. Les paramètres sont un ensemble de motifs P, un jeu de données $\mathcal{D}$ et un seuil ϵ. Ce dernier spécifie la différence minimale attendue entre deux pondérations issues d'itérations consécutives.

1. Initialiser le poids de chaque transaction t : $w_0[t] \leftarrow 1/|\mathcal{D}|$ et $i \leftarrow 0$
2. Définir le poids de chaque transaction t : $w_{i+1}[t] \leftarrow w_i[t] \times \frac{\sum_{t' \in \mathcal{D}} Supp_{w_i}(P_{\lhd t'}, \mathcal{D})/|\mathcal{D}|}{Supp_{w_i}(P_{\lhd t}, \mathcal{D})}$
3. Normaliser le poids $w_{i+1}[t]$ de chaque transaction t par la somme $\sum_{t \in \mathcal{D}} w_{i+1}[t]$
4. Recommancer à l'étape 2 en incrémentant i tant que $\sum_{t \in \mathcal{D}} |w_{i+1}[t] - w_i[t]|/|\mathcal{D}| \geq \epsilon$

Le fondement de SUPPORTBALANCE est de fixer le poids de chaque transaction de sorte qu'il soit inversement proportionnel à sa probabilité d'analyse. L'étape 2 calcule donc le nouveau poids w_{i+1} en multipliant w_i par le différentiel entre la couverture de t ($Supp_{w_i}(P_{\lhd t}, \mathcal{D})$) et la couverture moyenne ($\sum_{t' \in \mathcal{D}} Supp_{w_i}(P_{\lhd t'}, \mathcal{D})/|\mathcal{D}|$) conformément à l'équation 3.

La pondération issue de SUPPORTBALANCE permet de définir le *support équilibré* :

Définition 3 (Support équilibré) *Le support équilibré d'un motif $\varphi \in P$ dans $\mathcal{D}$ avec P et une erreur ϵ, dénoté par $\mathrm{BS}_\epsilon(\varphi, \mathcal{D}, P)$, est égal à son support pondéré $Supp_b(\varphi, \mathcal{D})$ où b est la pondération retournée par l'algorithme* SUPPORTBALANCE *avec les paramètres P, $\mathcal{D}$ et ϵ.*

Le support équilibré revalorise les motifs décrivant les transactions les moins décrites. Pour cette raison, le support équilibré de C, $\mathrm{BS}(C, \mathcal{D}, P) = 0.75$, est supérieur à son support usuel de 0.5 dans le jeu de données du tableau 1.

5 Expérimentations

L'objectif de cette étude empirique est de vérifier l'efficacité de l'algorithme d'équilibrage. Le tableau 3 reporte les résultats de SUPPORTBALANCE appliqué aux benchmarks de l'UCI[1] avec les motifs libres fréquents (Pasquier et al., 1999) où $minsupp = 0.05$, avec un seuil d'erreur $\epsilon = 10^{-5}$. La moyenne de la proportion d'analyse des transactions en utilisant $\mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, Supp)$ ou $\mathcal{A}_{\alpha,\delta}(\mathcal{D}, P, \mathrm{BS})$ sont respectivement indiquées dans les colonnes *Supp* et BS. De même, la variance est reportée dans les deux colonnes suivantes. Le gain précise le rapport entre la variance avec BS et la variance avec *Supp*. Enfin, le tableau indique l'écart moyen constaté entre les deux supports : $\sum_{\varphi \in P} |Supp(\varphi, \mathcal{D}) - \mathrm{BS}(\varphi)|/|P|$.

Le tableau 3 montre que SUPPORTBALANCE atteint son objectif puisqu'il diminue systématiquement la variance de la proportion d'analyse d'une transaction. La variance est au minimum diminuée de moitié et elle s'avère même divisée par plus de 10 dans 7 jeux de données. De plus, l'évolution du support classique à celui équilibré modifie profondément l'évaluation des motifs comme le montre l'écart moyen qui est toujours conséquent. Enfin, le nombre d'itérations nécessaire pour la convergence de l'algorithme est très variable suivant le jeu de données considéré.

6 Conclusion

Cet article a introduit le support équilibré qui induit une analyse équilibrée des motifs fréquents en considérant le modèle de l'analyse aléatoire. Nous avons aussi proposé l'algorithme

[1] `www.ics.uci.edu/~mlearn/MLRepository.html` et `users.info.unicaen.fr/~frioult/uci/uci.php`

				Moyenne		Variance			
Jeu de données	$\|\mathcal{D}\|$	$\|P\|$	Nbr. d'itér.	$Supp$	BS	$Supp$	BS	Gain	Ecart moyen
abalone	4177	2364	17	0.163	0.187	0.00643	0.00129	4.99	0.103
anneal	798	3290	29	0.204	0.237	0.0101	0.0021	4.78	0.0972
austral	690	13374	31	0.114	0.123	0.00181	0.000191	9.5	0.0624
breast	286	1823	42	0.171	0.163	0.00602	0.000355	17	0.079
cleve	303	10165	35	0.113	0.12	0.0044	0.000345	12.7	0.0629
cmc	1473	2632	23	0.148	0.144	0.00349	7.5e-05	46.6	0.0758
crx	690	17803	28	0.113	0.123	0.00213	0.000339	6.28	0.0675
german	1000	111047	17	0.0997	0.108	0.00331	0.000682	4.85	0.0665
glass	214	1920	57	0.126	0.138	0.00125	0.000151	8.25	0.0556
heart	270	13830	38	0.11	0.115	0.0024	0.000397	6.04	0.0547
iris	150	104	63	0.225	0.235	0.00281	0.000432	6.51	0.0694
page	941	2683	43	0.119	0.122	0.00125	0.000101	12.4	0.0496
pima	768	1035	51	0.128	0.136	0.000796	4.06e-05	19.6	0.0402
tic-tac-toe	958	1457	41	0.122	0.123	0.000503	2.42e-05	20.8	0.0365
vehicle	846	30480	30	0.0931	0.0993	0.00242	0.000249	9.72	0.0538
wine	178	6781	54	0.0976	0.105	0.00157	0.000133	11.8	0.0456
zoo	101	7057	53	0.245	0.241	0.00762	0.00149	5.13	0.114

TAB. 3 – *Benchmarks UCI avec les motifs libres fréquents (*$minsupp = 0.05$ *et* $\epsilon = 10^{-5}$*)*

SUPPORTBALANCE afin de calculer le support équilibré d'un ensemble de motifs. Les premiers résultats expérimentaux sont très prometteurs. Les perspectives d'utilisation du modèle de l'analyse aléatoire sont multiples. Il serait intéressant de construire un panel reflétant le jeu de données en choisissant les transactions les plus analysées suivant notre modèle. Nous souhaitons aussi améliorer la modélisation en tenant compte de l'ordre d'analyse des motifs.

Références

Agrawal, R. et R. Srikant (1994). Fast algorithms for mining association rules in large databases. In J. B. Bocca, M. Jarke, et C. Zaniolo (Eds.), *VLDB*, pp. 487–499. Morgan Kaufmann.

Brin, S. et L. Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks 30*(1-7), 107–117.

Fürnkranz, J. et A. Knobbe (2010). Guest editorial : Global modeling using local patterns. *Data Min. Knowl. Discov. 21*(1), 1–8.

Mannila, H. et H. Toivonen (1997). Levelwise search and borders of theories in knowledge discovery. *Data Min. Knowl. Discov. 1*(3), 241–258.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Inf. Syst. 24*(1), 25–46.

Summary

This paper proposes a method for evaluating the quality of patterns which benefits in advance from the analysis of patterns. We introduce the random analysis model of a pattern set. This model enables us to observe that the study of frequent patterns with the support leads to an unbalanced analysis of the dataset. In order to improve this analysis, we define the balanced support which corrects the usual support by weighting the transactions. We then propose an algorithm to compute these weights. Experimentations validate its efficiency.