Available online at https://www.asjp.cerist.dz/en/PresentationRevue/134

RIST

Information Processing at the Digital Age Journal

Extraction optimisée de règles d'association

Lamia Berkani[a,b,*], Lylia Betit[b], Yanis Chebahi[b]

[a]*Laboratoire en Intelligence Artificielle (LRIA), USTHB, Bab Ezzouar 16111, Alger, Algérie*
[b]*Département Informatique, USTHB, Bab Ezzouar 16111, Alger, Algérie*

Résumé

L'exploration de données, connue aussi sous différentes autres appellations dont la fouille de données, prospection de données ou le plus souvent data mining, a pour objet l'extraction de connaissances à partir de données volumineuses. L'utilisation des techniques de data mining consiste à convertir les données en connaissances appropriées pouvant être utiles à la prise de décision. Dans ce domaine, la recherche des règles d'association est une méthode répandue pour découvrir des relations ayant un intérêt pour la prise de décision en particulier lorsque les bases de données sont très volumineuses. Néanmoins, le défi majeur consiste à trouver les règles d'association les plus pertinentes. Dans cet article, nous proposons un framework dont l'objectif est la génération de règles d'association optimisées. Nous considérons trois approches: la première approche applique trois algorithmes de data mining: Apriori, Close et Charm; la second approche complète ces algorithmes en intégrant un algorithme génétique et la dernière approche applique d'abord l'algorithme génétique sur la base de données, dans un objectif de diversification suivi par des algorithmes de data mining. Notre objectif étant la génération des règles d'association les plus optimales en termes de pertinence, une nouvelle mesure de qualité des règles d'association générées a été proposée. Les résultats d'expérimentation démontrent l'apport de l'intégration des techniques de data mining avec les algorithmes génétiques.

Mots-clés: exploration de données; data mining ; règles d'association; optimisation ; algorithme génétique; mesure de pertinence.

1. Introduction

De nos jours, de grandes quantités de données sont collectées, manipulées et traitées par des organisations. Ces données sont stockées dans l'espoir d'être utiles à l'avenir (Amado et al., 2018). Selon ces auteurs, de nouveaux défis sont soulevés en ce qui concerne la gestion de ces données et l'extraction des connaissances appropriées pour appuyer les décisions. Le processus d'aide à la décision constitue un support important pour

* Corresponding author: lberkani@usthb.dz

les spécialistes du marketing. Il fournit des conseils pour répondre à des questions critiques telles que: quel est le produit le plus approprié pour une catégorie de consommateurs, comment faire la publicité d'un tel produit sur le marché, à quels moments et à quel prix et soutenu par quel type d'actions promotionnelles et publicitaires (Amado et al., 2018). D'autre part, les chefs d'entreprise sont confrontés au défi de réussir à exploiter et d'améliorer continuellement leur organisation. Ainsi, leur objectif principal est de prendre les meilleures décisions possibles. Les approches traditionnelles en matière de prise de décision reposent souvent sur l'expérience du décideur. Cependant, ces approches, en plus du fait qu'elles ne peuvent pas garantir un meilleur résultat comportent de nombreuses limitations notamment en termes de temps et de coût (Goienetxea et al., 2015; Goienetxea et al., 2017). Afin de faciliter ce processus, la tendance actuelle consiste à exploiter les techniques de data mining (DM), i.e. d'exploration de données à savoir un ensemble de techniques et de méthodologies de calcul visant à extraire des connaissances à partir d'un grand nombre de données (Urso et al., 2018).

Dans le domaine du data mining, la recherche des règles d'association est une méthode populaire étudiée d'une manière approfondie dont le but est de découvrir des relations ayant un intérêt pour la prise de décision en particulier lorsque les bases de données sont très volumineuses. En se basant sur le concept de relations fortes, Agrawal et al. (1994) présentent des règles d'association dont le but est de découvrir des similitudes entre des produits. Les règles d'association sont employées aujourd'hui non seulement comme base pour des décisions marketing (des promotions ou des placements bien choisis pour les produits associés…), mais aussi dans plusieurs autres domaines incluant celui de la fouille du web, détection d'intrusion , ou encore la bio-informatique.

Nous nous intéressons dans cet article à l'application de techniques de data mining afin de permettre une meilleure prise de décision. Nous proposons un framework d'aide à la décision utilisant trois algorithmes différents de data mining: Apriori, Close et Charm. L'objectif étant la génération d'un ensemble de règles d'association pour l'extraction de connaissances. De plus, afin d'avoir les solutions les plus optimales, nous avons intégré dans notre approche les algorithmes génétiques. L'un des résultats les plus importants de notre contribution est la proposition d'une nouvelle mesure de pertinence pour filtrer les règles d'association générées.

Le reste de cet article sera organisé comme suit: La section 2 présente un état de l'art sur l'utilisation du data mining et des techniques d'optimisation. La section 3 propose un framework d'aide à la décision basé sur les algorithmes de data mining et des algorithmes génétiques. Les résultats d'expérimentation sont décrits dans la section 4. Enfin, la section 5 présente la conclusion et quelques perspectives.

2. Etat de l'art

Cette section présente les concepts de base concernant les algorithmes de data mining et les règles d'association, ainsi que des travaux liés à l'utilisation des algorithmes de data mining et d'optimisation.

2.1. Algorithmes de Data Mining

Le Data Mining est défini comme l'ensemble des techniques de calcul visant à extraire des connaissances d'une grande quantité de données (Urso et al., 2018). Il est considéré comme un champ multidisciplinaire, intégrant des approches et des techniques issues de diverses disciplines telles que l'apprentissage automatique, les statistiques, l'intelligence artificielle et la gestion de bases de données. Plusieurs algorithmes ont été proposés (Wu et al., 2008), notamment les algorithmes Apriori, Close et Charm. L'algorithme Apriori (Agrawal and Sikrant, 1994) sert à reconnaitre des propriétés qui reviennent fréquemment dans un ensemble

de données et d'en déduire une catégorisation. Cet algorithme considère deux paramètres: minsupp, l'indice de support minimum donné, et minconf, l'indice de confiance donné et s'exécute en deux étapes: (1) génération de tous les itemsets fréquents; et (2) génération de toutes les règles d'associations de confiance à partir des itemsets fréquents. L'algorithme Close (Pei et al., 2000) repose sur l'extraction de générateurs d'ensembles de mots fermés fréquents. L'algorithme Charm (Zaki et Hsiao, 2002) explore simultanément l'espace des itemsets et l'espace tidset à l'aide de l'arborescence IT-tree, contrairement aux méthodes précédentes qui exploitaient généralement uniquement l'espace des itemsets.

2.2. Règles d'association

Une règle d'association est une méthode répandue pour découvrir des relations intéressantes entre des variables dans de grandes bases de données (Indira and Kanmani, 2011). Il s'agit d'étudier la fréquence des éléments réunis dans des bases de données transactionnelles en se basant sur deux seuils: (1) le support (s), identifie les ensembles d'éléments fréquents (itemsets fréquents); et (2) la confiance (c), est la probabilité conditionnelle qu'un élément apparaisse dans une transaction lorsqu'un autre élément apparaît, et est utilisée pour identifier des règles d'association.

Les règles d'association découvertes suivent le format suivant: $X \rightarrow Y\ [s,\ c]$, où X et Y sont des conjonctions de paires d'attribut-valeur, « s » est la probabilité que X et Y apparaissent ensemble dans une transaction et « c » est la probabilité conditionnelle que Y apparaît dans une transaction lorsque X est présent.

2.3. Travaux liés

La revue de la littérature montre un nombre considérable de travaux utilisant les techniques de data mining dans les organisations: pour identifier les indicateurs de performance clés pertinents (Peral et al., 2017), détecter les retraits financiers (Dutta et al., 2017), améliorer la prévision du taux de désabonnement des clients dans le secteur des télécommunications (Keramati et al., 2014), analyser les données des utilisateurs de lignes fixes pour établir un système de prévision des retards de paiements (Chen et al., 2013).

L'étude de ces travaux nous a montré que les algorithmes de data mining sont souvent combinés avec d'autres techniques. Dans (Chen et al., 2013) par exemple, les auteurs ont adopté une stratégie d'exploration de données dirigée par le domaine et ont utilisé des règles d'association, le clustering et les arbres de décision. En outre, l'étude d'autres travaux, nous a montré que certains auteurs ont utilisé des techniques d'optimisation pour l'exploration de données. Par exemple, Yu et al. (2005) ont utilisé les algorithmes génétiques pour extraire des stocks de haute qualité à des fins d'investissement. Agarwal (2015) a adapté les algorithmes génétiques aux algorithmes de datamining proposant une stratégie de recherche pour trouver des connaissances précises et compréhensibles dans une base de données volumineuse pouvant être considérée comme un espace de recherche.

Cependant, à notre connaissance, nous n'avons pas trouvé de travaux combinant des techniques de datamining avec des algorithmes d'optimisation pour la génération de règles d'association. Par conséquent, nous tentons dans ce travail de combiner ces techniques pour obtenir des solutions optimisées pour une meilleure prise de décision.

3. Proposition d'un framework d'aide à la décision

Nous considérons trois approches différentes pour la génération de règles d'associations (RAs): (1) l'application des algorithmes de data mining ; (2) l'application des algorithmes génétiques à base d'algorithme de data mining et (3) l'application des algorithmes de data mining à base d'algorithme génétique.

3.1. Approche 1- Application d'algorithmes de data mining

Dans cette approche, trois algorithmes d'exploration de données sont développés: A priori, Close et Charm, conformément au processus classique d'extraction de RAs composé de deux étapes: la recherche de données et la génération de RAs:

- *La recherche de données:* concerne l'extraction d'itemsets fréquents. Cette étape est la plus coûteuse en raison de la taille et du nombre nécessaire de balayages complets de la base de données.
- *La génération de RAs:* une fois l'ensemble d'itemsets fréquents obtenu, les RAs ayant un degré de confiance supérieure au seuil de confiance minimal (MinConf), fourni par l'utilisateur, sont générées.

3.2. Approche 2- Application d'un algorithme génétique à base d'algorithmes de data mining

Pour la génération d'une RA, le critère du seuil minimal de confiance n'est pas suffisant, car nous recherchons également la pertinence de la RA. Par conséquent, nous avons proposé une deuxième approche qui combine la première avec un algorithme génétique (voir Fig. 1).

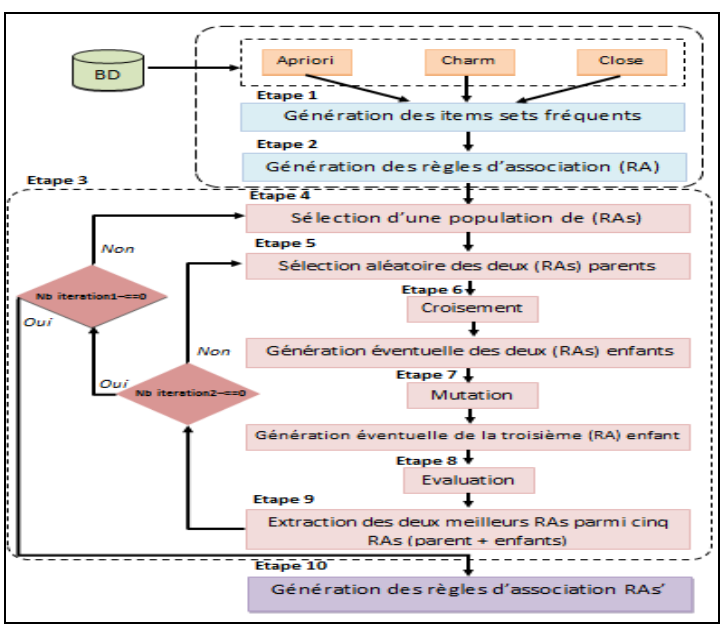
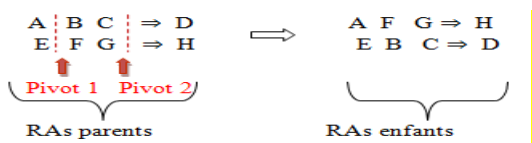


Fig. 1. Génération de RAs utilisant l'AG basé sur les algorithmes de DM.

La Fig. 1 illustre l'enchainement des différentes étapes de l'approche 2, où *nbIteration1* représente le nombre de fois de sélection d'un ensemble de RAs et *nbIteration2* est le nombre de fois de sélection des deux RAs parents à partir de l'ensemble sélectionné. Les étapes de cette approche sont décrites comme suit :

- *Étape 1:* correspond à la recherche des données.
- *Étape 2:* représente la génération des RAs, comme dans l'approche 1.
- *Étape 3:* concerne l'application de l'algorithme génétique incluant les différentes étapes de fonctionnement de cet algorithme appliqué sur l'ensemble des règles d'association extraites RA (résultat de l'approche 1).
- *Étape 4:* sélection aléatoire de *n* règles d'association, afin de former une population initiale de RAs. Nous notons par « *PS* » la Population Sélectionnée où *n* est un nombre entier inférieur strictement au nombre total des RAs générées dans l'étape 2, et qui sera défini par l'utilisateur.
- *Étape 5:* est la sélection aléatoire de deux RAs à partir de la *PS*, appelées "parents". Les parents ainsi sélectionnés seront soumis aux opérations génétiques suivantes : la première étant l'opération de croisement (étape 6) ensuite la mutation (étape7) et enfin l'évaluation (étape8).
- *Étape 6:* concerne l'opération de croisement. Dans cette étape, nous proposons d'effectuer un simple enjambement ou un double enjambement.
    - *Simple enjambement:* croisement selon un seul site (1 pivot), comme illustré par l'exemple suivant :

    - *Double enjambement :* croisement sur deux sites (2 pivots), comme illustré par l'exemple suivant :



Pour une RA "*R*" donnée avec "*n*" items, l'ensemble des sites d'un pivot est égal à "*n*". Ainsi, avec *n* positions, nous aurons un ensemble fini de *n* règles différentes. Par ailleurs, l'ensemble des sites de déplacement de deux pivots sera calculé par la probabilité combinatoire comme suit:

$$C_n^2 = \frac{n(n-1)}{2} \qquad (1)$$

Ce qui donne lieu à n (n-1) /2 règles différentes.

- *Étape 7:* concerne la mutation qui consiste à choisir aléatoirement une RA parmi les règles parents et enfants, puis positionner aléatoirement le site de mutation sur un élément de la règle choisie afin de le muter. Si la RA générée n'est pas différente des RAs parents et enfants, donc elle n'apporte pas de nouvelles informations, dans ce cas aucune considération n'est donnée à cette dernière.

- *Étape 8:* consiste à effectuer une évaluation des règles selon deux mesures d'intérêt différentes, dont une seule sera choisie par l'utilisateur:
    - *Lift:* il s'agit de calculer pour les cinq règles d'association obtenues leur lift selon la formule déjà présentée (deux RAs parents et trois éventuelles RAs enfants).
    - *Relation de dominance:* consiste à calculer toutes les mesures choisies, i.e. support, confiance et lift. Sur la base du principe de la relation de dominance, nous avons proposé un nouvel indicateur de qualité appelé "Dominance pondérée" (DominanceW- Weighted Dominance). Comme son nom l'indique, la dominance pondérée est fondée sur un ensemble de poids qui seront affectés aux mesures choisies, tel que ces poids représentent les pondérations d'influence des mesures. Son principe consiste à comparer les RAs deux à deux, en calculant la somme de la différence entre les valeurs des deux règles pour chaque mesure. Sa formule est donnée comme suit :

$$Dominance_W(R_1, R_2) = \sum_{i=1}^{i=3} ((M_i(R_1) - M_i(R_2)) \times P_i) \quad (2)$$

où:

R_1 et R_2 sont deux RAs;
M_i : est la i^{ème} mesure choisie; et
P_i : est le poids affecté à chaque i^{ème} mesure, avec $\sum p_i = 1$.
Le tableau 1 décrit les interprétations des valeurs de $Dominance_W$:

Tableau 1. Description des valeurs de Dominance pondérée

Valeur	Interprétation
> 0	La règle R_1 est dominante
< 0	La règle R_2 est dominante
= 0	Comparer la $Dominance_W$ de toutes les mesures ayant des pondérations importantes.

Étape 9: concerne l'extraction des deux meilleurs RAs. Si l'utilisateur choisit le lift comme mesure alors la valeur la plus élevée détermine ces deux règles. Par ailleurs, si l'utilisateur choisit la relation de dominance, ces deux règles (i.e. règles dominantes) seront déterminées par la comparaison entre les mesures et le calcul de la moyenne pondérée pour toutes les RAs. Ces deux règles, une fois déterminées, seront ajoutées à *PS* à la place des deux règles parents. Une boucle est définie pour retourner à l'étape 5 (parcourir les étapes 5 à 9 jusqu'à atteindre *nbIteration2*). Une nouvelle séquence est lancée par une nouvelle boucle (pour exécuter les étapes 4 à 9), après intégration de *PS* dans l'ensemble initial de RAs générées par la première approche. Pour chaque nouvelle itération, toutes les RAs sélectionnées à l'étape 4 de l'itération précédente seront supprimées. Cette boucle s'arrête lorsque *nbIteration1* = 0. Ces deux boucles ont pour rôle d'exécuter l'algorithme génétique autant de fois que l'utilisateur le souhaite afin d'accroître la pertinence des RAs.

- *Étape 10:* à cette étape, un nouvel ensemble R' de règles d'association est déterminé.

3.3. Approche 3- Application des algorithmes de data mining à base d'algorithme génétique

Cette approche applique les algorithmes de data mining sur les résultats obtenus par l'application d'un algorithme génétique sur la base de données, comme présenté par la figure suivante. L'objectif de l'application de l'algorithme génétique étant de prévoir d'autres types de transactions qui n'existent pas dans la base initiale mais que les étapes de mutation et de croisement de l'algorithme génétique pourraient générer (ce qui correspond au caractère biologique du principe de l'hérédité sur lequel est basé cet algorithme):

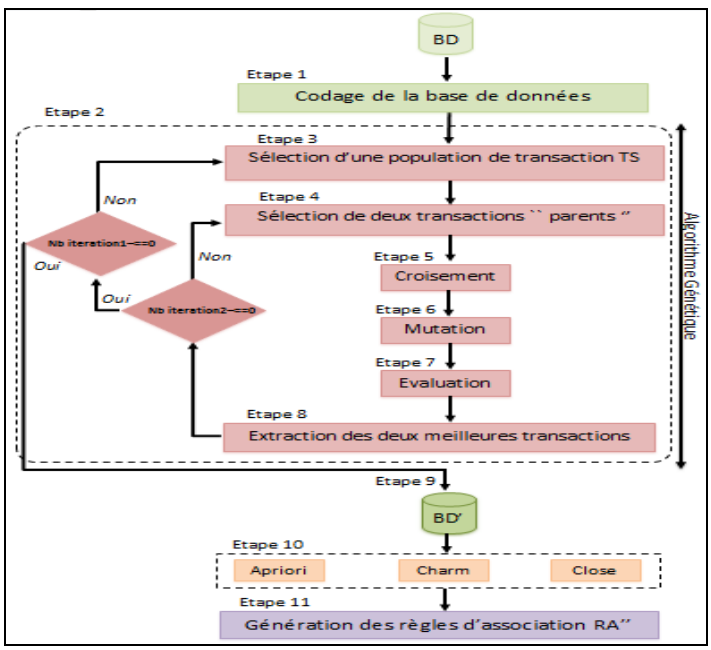
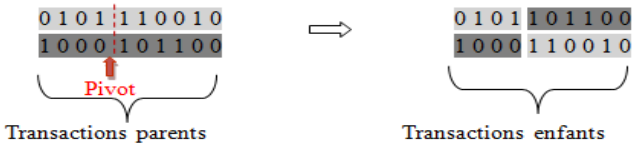


Fig. 2. Génération de RAs utilisant les algorithmes de DM basé sur un algorithme génétique

Les étapes de l'approche 3 sont décrites comme suit:

- *Étape 1:* consiste en un encodage binaire des transactions par des chaînes de bits 0 si l'attribut n'existe pas dans la transaction et 1 sinon.
- *Étape 2:* concerne l'application d'un algorithme génétique (allant de l'étape 3 jusqu'à l'étape 8) sur la base de données transactionnelles initiale BD, codée en binaire.
- *Étape 3:* représente la sélection aléatoire de *n* transactions, afin de générer une population, appelée *TS* (Transactions Sélectionnées).
- Étape 4: sélection aléatoire de deux transactions à partir de TS, appelées "parents".

- *Étape5:* pour cette étape nous avons choisi d'effectuer un simple enjambement, afin de croiser les transactions parents et générer deux nouvelles transactions enfants. L'exemple suivant décrit le simple enjambement sur une BD transactionnelle contenant dix attributs :



- *Étape 6:* consiste à choisir aléatoirement une transaction parmi les transactions parents et enfants, puis fixer aléatoirement un bit et le muter (si le bit est égal à 1 alors le mettre à 0 et inversement s'il est égal à 0 alors le mettre à 1).
- *Étape 7:* évaluer les transactions parents et enfants, afin de pouvoir choisir les meilleures selon leurs fitness. Pour cela nous avons proposé une fonction qui consiste à calculer la moyenne des supports *MoySupp* des attributs de chaque transaction.

$$MoySupp(t) = \frac{\sum_{i=1}^{i=n} s_i}{n} \qquad (3)$$

où :

t est une transaction ;
S_i est le support de l'i^{ème} attribut de la transaction *t*; et
n est le nombre total d'attributs existants dans *t*.

Étape 8: permet l'extraction des deux meilleures transactions. Une fois les transactions parents et enfants évaluées, notre choix s'est porté sur deux transactions ayant la valeur la plus petites de *MoySupp*. L'objectif étant de changer les supports des attributs de la BD et donc diversifier les itemsets fréquents dans cette base. À chaque itération et après avoir ajouté les deux transactions dans la *TS* à la place des deux transactions parents, il s'agit de relancer l'exécution de l'étape 4 jusqu'à l'atteinte du nombre de fois de sélection des transactions parents souhaité par l'utilisateur (*nbIteration2*).
Ensuite, une autre itération de la boucle principale est déclenchée à partir de l'étape 3 où une nouvelle *TS* est générée (après avoir ajouté la *TS* précédente dans la BD à la place de l'ensemble de transactions sélectionnées à l'étape 3 de l'itération précédente). Cette boucle s'arrête lorsque la condition d'arrêt est vérifiée (*nbIteration1* = 0).

Étape 9: l'application de l'algorithme génétique sur la BD, donne comme résultat une nouvelle base de données BD', contenant exactement le même nombre de transactions que la base initiale. Finalement, dans le but de générer les règles d'association RA", l'approche 1 sera appliquée sur la BD' (les étapes 10 et 11).

4. Expérimentation

Nous avons développé un système de génération de RAs basé sur les trois approches décrites dans la section précédente. Afin d'évaluer notre solution, nous avons effectué une série d'expérimentations à l'aide de

la BD Epicerie (Grocery), collectée à partir du site : https://www.analyticsvidhya.com/ et incluant 9835 transactions.

4.1. Temps de réponse des algorithmes de DM

Nous avons défini Minsup à 0,01 et Minconf à 0,03, le nombre d'attributs à 169, en faisant varier le nombre de transactions dans la BD. La Fig. 3 illustre l'évaluation des temps de réponse des algorithmes de DM en fonction du nombre de transactions.

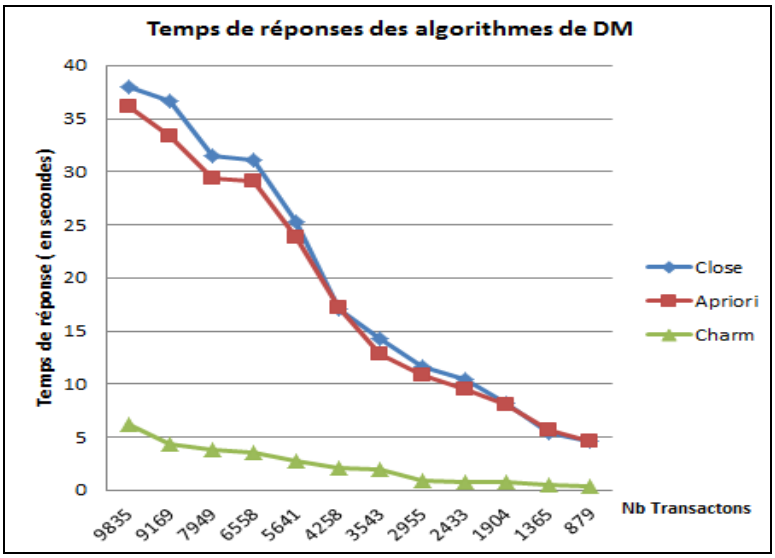


Fig. 3. Evaluation du temps de réponse des algorithmes de DM.

Les résultats montrent que l'algorithme Charm est plus rapide grâce à son principe de fonctionnement et à sa structure arborescente. Les deux autres algorithmes ont également donné des résultats satisfaisants compte tenu du grand nombre de balayages de la BD imposés par le principe de leur fonctionnement. Ces résultats ont été obtenus grâce à notre utilisation de Hashmaps pour la mise en œuvre de ces algorithmes.

Nous avons répété la même évaluation avec une variation de mimimum support (MinSup) et minimum confiance (MinConf), en utilisant une base de données ayant 1 000 transactions et 169 attributs. Le tableau 2 présente les résultats obtenus:

Tableau 2. Evaluation du temps de réponse avec variation de MinConf et MinSup

MinConf (%)	Minsup (%)	Apriori	Close	Charm
1	0.3	244.5	293.13	7.76
	0.7	6.86	61.31	5.32
	1.0	34.52	37.27	5.21
	3.0	6.24	6.36	4.46
	5.0	3.14	2.71	3.00
3	0.3	258.00	299.00	7.29
	0.7	101.40	110.24	6.02
	1.0	37.7	36.81	4.04
	3.0	5.59	5.69	3.67
	5.0	3.10	2.58	2.42
5	0.3	242.14	276.10	6.29
	0.7	60.69	63.51	5.95
	1.0	32.00	36.12	4.42
	3.0	5.13	5.28	4.54
	5.0	2.40	2.37	2.38

Nous notons l'existence d'une relation inverse entre les valeurs de MinSup et MinConf en ce qui concerne les temps d'exécution des algorithmes. Étant donné que MinSup et MinConf sont élevés, les temps de réponse des algorithmes diminuent, en raison de la restriction de l'espace de recherche des algorithmes.

4.2. Comparaison des mesures de performance

Afin de comparer la pertinence des trois mesures de qualité (lift, dominance et dominance pondérée), nous avons considéré la deuxième approche (algorithme génétique à base de data mining) en faisant varier la taille de la BD et en définissant les paramètres suivants: NbIteration1 = 1000, NbIteration2 = 90, MinSup = 1% et MinConf = 3%. Nous pouvons constater que les valeurs des trois mesures sont convergentes, avec une meilleure performance obtenue avec la mesure de dominance pondérée (voir Tableau 3):

Tableau 3. Comparaison entre Lift, Dominance et DominanceW

BD	# Transactions	# Attributs	Lift	Dominance	Dominance w
BD 1	9835	169	1,79	1,70	1,75
BD 2	8000	169	1,74	1,71	1,74
BD 3	5500	168	1,84	1,81	1,89
BD4	9639	91	2,4	2,38	2,47
BD 5	6000	91	2,39	2,36	2,43

4.3. Évaluation du nombre de RAs générées

L'objectif de cette évaluation est de comparer le nombre de règles générées par les trois approches en faisant varier la taille de la BD. Nous avons utilisé l'algorithme Charm (qui offre de meilleures performances en termes de temps de réponse) et défini les paramètres suivants: MinSup = 1%, MinConf = 3%, nbIteration 1 = 200, nbIteration2 = 25. En ce qui concerne l'algorithme génétique, nous avons choisi une combinaison commune: 1 pivot (pour le croisement) et le lift (pour l'évaluation). Les résultats sont présentés dans le Tableau 4:

Tableau4. Evaluation du nombre de RAs générées

Approches	BD	# Transactions	# Attributs	# RAs
Approche 1	BD1	9835	169	599
	BD2	9639	91	745
	BD3	5000	169	595
	BD4	6285	91	766
Approche 2	BD1	9835	169	590
	BD2	9639	91	736
	BD3	5000	169	586
	BD4	6285	91	757
Approche 3	BD1	9835	169	1022
	BD2	9639	91	3025
	BD3	5000	169	1002
	BD4	6285	91	3093

La Fig. 4 montre que la troisième approche a généré plus de RAs, avec la variation de la taille de la BD. Ce résultat est justifié par le fait que cette approche vise à améliorer la BD en augmentant le nombre d'itemsets fréquents (ce qui a par conséquent augmenté le nombre de règles générées).

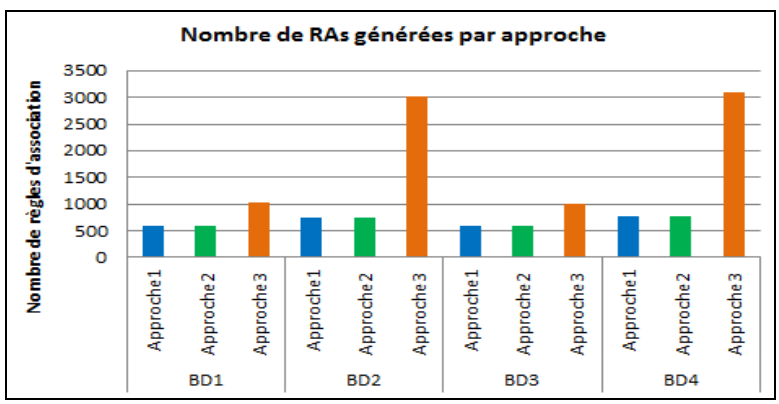


Fig. 4. Comparaison par rapport au nombre de RAs générées.

4.4. Evaluation de pertinence des RAs générées par l'approche 2

L'objectif de cette évaluation est de déterminer la meilleure combinaison des différents paramètres de l'approche 2, en termes de pertinence de l'ensemble des règles générées. Nous avons choisi une BD avec 9835 transactions et 168 attributs, et nous avons modifié les conditions d'arrêt de l'AG (nbIteration1, nbIteration2). Le reste des paramètres est fixé comme suit: MinSup = 0.3% et MinConf= 3%. Nous avons également varié les combinaisons de mesures de croisement et d'évaluation en considérant quatre évaluations différentes: un croisement de 1 pivot ou de 2 pivots et les mesures Lift ou DominanceW (c.-à-d. 1 pivot et Lift; 1 pivot et DominanceW; 2 pivots et Lift; 2 pivots et DominanceW). Les résultats sont donnés par le Tableau 5:

Nous avons considéré dans cette évaluation la moyenne lift, Moy_{Lift}, qui est calculée selon la formule suivante, sachant que l'ensemble de RAs ayant la plus grande valeur de Moy_{Lift} est considéré comme pertinent:

$$Moy_{Lift} = \sum_{i=1}^{i=n} \frac{Lift(R_i)}{n} \qquad (4)$$

avec:

R_i est la i^{ème} règle avec Lift > 1; et
n est le nombre de règles ayant un Lift > 1.

Comme les algorithmes génétiques des approches 2 et 3 se basent sur l'aléatoire, nous avons utilisé la Moy_{Lift} de plusieurs exécutions, puis nous avons calculé la moyenne des résultats obtenus.

Tableau 5. Evaluation de pertinence des RAs générées par l'Approche 2

Croisement & Evaluation	NbIteration1	NbIteration2	Moyenne LIFT
1 pivot et Lift	300	40	1,74
	1000	90	1,73
	500	10	1,71
	2000	200	1,72
1 pivot et Dominancew	300	40	1,87
	1000	90	1,88
	500	10	1,74
	2000	200	1,75
2 pivots et Lift	300	40	1,73
	1000	90	1,73
	500	10	1,72
	2000	200	1,71
2 pivots et Dominencew	300	40	1,96
	1000	90	1,95
	500	10	1,77
	2000	200	1,75

4.5. Evaluation de pertinence des RAs générées par les trois approches

Cette évaluation vise à comparer les trois approches en termes de qualité des ensembles de règles d'association RA, RA' et RA" générées respectivement par les approches 1, 2 et 3. En utilisant les mêmes paramètres que ceux de l'évaluation précédente, nous avons obtenu les résultats suivants:

Tableau 6. Comparaison de pertinence des RAs générées

Approches	BD	# Transactions	# Attributs	MOY Lift
Approche 1	BD1	9835	169	1,67
	BD2	9639	91	2,32
	BD3	5000	169	1,70
	BD4	6285	91	2,39
Approche 2	BD1	9835	169	1,73
	BD2	9639	91	2,37
	BD3	5000	169	1,77
	BD4	6285	91	2,42
Approche 3	BD1	9835	169	3,89
	BD2	9639	91	1,85
	BD3	5000	169	3,76
	BD4	6285	91	1,84

Nous pouvons facilement remarquer que la troisième approche est plus pertinente avec les jeux de données 1 et 3, mais moins pertinente avec les jeux de données 2 et 4. Cela est dû à la diminution du nombre d'attributs dans ces deux derniers jeux de données. De plus, les règles générées par la deuxième approche sont plus pertinentes que celles de la première. Cela est dû au fait que l'approche 2 améliore les résultats obtenus par l'approche 1 en utilisant l'algorithme génétique.

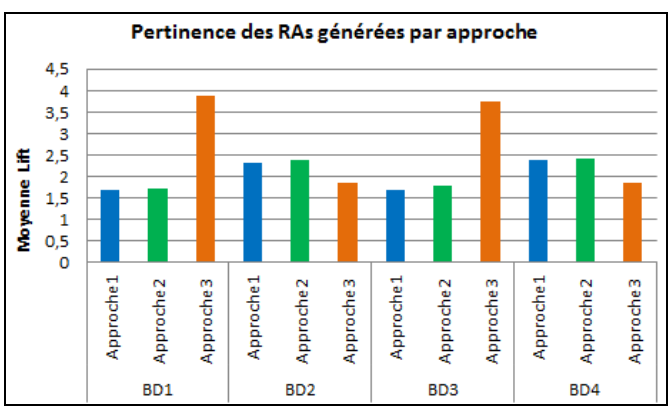


Fig. 5. Comparaison de pertinence des RAs générées

5. Conclusion

Le but de cette recherche est d'optimiser les règles d'association générées. Trois approches différentes ont été considérées : la première approche applique différents algorithmes de data mining (Apriori, Close et Charm), la seconde approche génère les meilleures règles d'association à l'aide d'un algorithme génétique basé sur les algorithmes de data mining et la troisième approche génère l'ensemble de règles d'association à l'aide d'algorithmes de data mining basés sur un algorithme génétique. Les résultats d'expérimentation réalisés sur la base Epicerie ont montré que l'intégration des algorithmes génétiques et de data mining ont donné de meilleures performances en termes de nombre et de pertinence des règles d'association générées.

Comme perspectives à ce travail, nous envisageons d'utiliser d'autres algorithmes d'optimisation et de data mining. Nous prévoyons également d'appliquer cette recherche dans le domaine de télécommunications mobiles. Ce qui permettrait d'avoir une meilleure analyse des clients et des ventes de ces entreprises et donc offrir aux responsables le moyen de prédire les offres adéquates et augmenter en conséquence leurs chiffres d'affaires.

References

Agarwal, R., 2015. Genetic Algorithm in Data Mining, International Journal of Advanced Research in Computer Science and Software Engineering, vol. No 9, pp. 631-634.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, Santiago, Chile, pp. 487-499

Amado, A., Cortez, P., Ritac, P., Moro, S., 2018. Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics, vol. 24, pp. 1–7, 2018.

Chen, C-H., Chiang, R-D, Wu, T-F. Chu, H-C. , 2013. A combined mining-based framework for predicting telecommunications customer payment behaviors, Expert Systems with Applications, vol. 40, No 16, pp. 6561-6569.

Dutta, I., Dutta, S., Raahem, B., 2017. Detecting financial restatements using data mining techniques", Expert Systems with Applications, vol 90, , pp. 374-393.

Goienetxea Uriarte, A., Ruiz Zúñiga, E., Urenda Moris, M., Ng, A.H.C., 2017. How can decision makers be supported in the improvement of an emergency department?" A simulation, Optimization and DM Approach Operations Research for Health Care , vol. 15, pp.102-122.

Goienetxea Uriarte, A., Urenda Moris, M.., Ng, A.H.C., Oscarsson, J., 2015. Lean, simulation and optimization: a win-win combination". In: W.K.V.C.L. Yilmaz, I. Moon, T.M.K. Roeder, C. Macal, M.D. Rossetti (Eds.), Winter Simulation Conference, IEEE, Huntington Beach, California, USA, pp. 2227–2238.

Indira, K., Kanmani, S., 2011. Association Rule Mining Using Genetic Algorithm: The Role of Estimation Parameters. In: A. Abraham et al. (Eds.): ACC 2011, Part I, CCIS 190, Springer-Verlag Berlin Heidelberg, pp. 639–648.

Keramati, R., Jafari-Marandi, M., Aliannejadi, I., Ahmadian, M., Mozaffari, U., Abbasi, U., 2014. Improved churn prediction in telecommunication industry using data mining techniques, Applied Soft Computing, vol. 24, pp. 994-1012.

Pei, J., Han, J., Mao, R., 2000. Closet: An effcient algorithm for mining frequent closed itemsets. In: SIGMOD International Workshop on DM and Knowledge Discovery.

Peral, J., Maté, A., Marco, M., 2017. Application of Data Mining techniques to identify relevant Key Performance Indicators", Computer Standards and Interfaces, vol 54, pp. 76-85.

Urso, A., Fiannaca, A., La Rosa, M., Ravì, V., Rizzo, R., 2018. Data Mining: Classification and Prediction, Reference Module in Life Sciences.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J. , Yang, Q., Motoda, H.,. McLachlan, G.J, Ng, A., Liu, B.,. Yu, P.S, Zhou, Z., Steinbach, M., Hand D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. Survey Paper, Knowledge Information System, vol. 14, pp. 1-37.

Yu, L., Lai, K.K., Wang, S., 2005. Mining Valuable Stocks with Genetic Optimization Algorithm, Optimization-based Data Mining Techniques with Applications, Proceedings of a Workshop in IEEE International Conference on Data Mining, Houston, USA.

Zaki M.J., Hsiao, C.J., 2002. Charm : An efficient algorithm for closed itemset mining. In Robert L. Grossman, Jiawei Han, Vipin Kumar, Heikki Mannila et Rajeev Motwani, éditeurs : SDM. SIAM, http://epubs.siam.org/doi/pdf/10.1137/1.9781611972726.27.