Modèle décisionnel basé sur la qualité des données pour sélectionner les règles d'associations légitimement intéressantes

Laure Berti-Équille
IRISA, Campus Universitaire de Beaulieu,
35042 Rennes, France
berti@irisa.fr

Résumé. Dans cet article nous proposons d'exploiter des mesures décrivant la qualité des données pour définir la qualité des règles d'associations résultant d'un processus de fouille. Nous proposons un modèle décisionnel probabiliste basé sur le coût de la sélection de règles légitimement, potentiellement intéressantes ou inintéressantes si la qualité des données à l'origine de leur calcul est bonne, moyenne ou douteuse. Les expériences sur les données de KDD-CUP-98 montrent que les 10 meilleures règles sélectionnées d'après leurs mesures de support et confiance ne sont intéressantes que dans le cas où la qualité de leurs données est correcte voire améliorée.

1 Introduction

La qualité des règles d'association est généralement évaluée par des mesures d'intérêt (classiquement le support et la confiance) et de nombreuses autres mesures ont été proposées (Tan *et al.*, 2002). Mais, on peut légitimement se demander quel est l'intérêt de telles règles, notées LHS → RHS, si 30 % des données de LHS sont obsolètes, 20% des données de RHS sont imprécises, et 15% des données de LHS proviennent d'une source réputée peu fiable. La thèse défendue dans cet article est que les mesures d'intérêt pour la découverte de règles d'associations ne sont pas autosuffisantes pour représenter effectivement la qualité des règles. Des mesures décrivant la qualité des données à partir desquelles sont calculées les règles doivent être intégrées au processus de découverte, de même que le coût d'une décision de choisir (ou non) ces règles « supposées intéressantes » doit être également considéré. Ceci a motivé donc nos travaux que nous formalisons dans les sections suivantes.

2 Caractérisation de la qualité des règles d'association à partir de la qualité des données d'origine

Soit I un ensemble d'items. Une règle d'association R est une implication de la forme: $LHS \rightarrow RHS$ où $LHS \subseteq I$, $RHS \subseteq I$ et $LHS \cap RHS = \varnothing$. LHS et RHS sont des conjonctions de variables telles que l'extension de LHS est : $g(LHS) = x_1 \wedge x_2 \wedge \ldots \wedge x_n$ et l'extension de Y est $g(RHS) = y_1 \wedge y_2 \wedge \ldots \wedge y_{n'}$. Soit j ($j=1, 2, \ldots, k$) une dimension décrivant un aspect de la qualité des données (*i.e.*, complétude, fraîcheur, précision, cohérence, crédibilité, etc.). Soit $q_j(I_i) \in [min_{ij}, max_{ij}]$ le score de la dimension de qualité j pour le sous-ensemble de données I_i

($I_i \subseteq I$). Pour chaque itemset I_i, le vecteur composé des scores sur toutes les dimensions de qualité (normalisées sur [0,1]) est appelé vecteur qualité et noté $q(I_i)$ dans l'espace qualité Q de tous les vecteurs qualité possibles. Nous définissons la qualité d'une règle d'association R avec une fonction de fusion notée "o_j" dont la sémantique est spécifique à la dimension de qualité j considérée. Cette fonction permet de fusionner chaque composante des vecteurs qualité correspondants aux ensembles de données présents dans les partie gauche et droite de la règle. La qualité de la règle R est donc un vecteur k-dimensionnel tel que:

$$Quality(R) = \begin{pmatrix} q_1(R) \\ q_2(R) \\ \vdots \\ q_k(R) \end{pmatrix} = \begin{pmatrix} q_1(LHS) \circ_1 q_1(RHS) \\ q_2(LHS) \circ_2 q_2(RHS) \\ \vdots \\ q_k(LHS) \circ_k q_k(RHS) \end{pmatrix}$$

$$= \begin{pmatrix} q_1(x_1) \circ_1 q_1(x_2) \circ_1 \cdots \circ_1 q_1(x_n) \circ_1 q_1(y_1) \circ_1 q_1(y_2) \circ_1 \cdots \circ_1 q_1(y_{n'}) \\ q_2(x_1) \circ_2 q_2(x_2) \circ_2 \cdots \circ_2 q_2(x_n) \circ_2 q_2(y_1) \circ_2 q_2(y_2) \circ_2 \cdots \circ_2 q_2(y_{n'}) \\ \vdots \\ q_k(x_1) \circ_k q_k(x_2) \circ_k \cdots \circ_k q_k(x_n) \circ_k q_k(y_1) \circ_k q_k(y_2) \circ_k \cdots \circ_k q_k(y_{n'}) \end{pmatrix} \quad (1)$$

On peut alors définir la qualité moyenne de la règle R notée $\overline{q}(R)$ par une somme pondérée de chaque dimension des vecteurs qualité des jeux de données composant la règle : $\overline{q}(R) = \sum_{j=1}^{k} w_j . q_j(R)$ avec $\sum_{j=1}^{k} w_j = 1 \ \forall j = 1,2,\ldots k$ (2)

Le Tableau 1 présente plusieurs exemples de définition ainsi que la sémantique que l'on peut donner à la fonction de fusion pour combiner les scores de qualité sur la dimension considérée pour deux ensembles de données x et y composant la règle.

j	DIMENSION	FONCTION DE FUSION "o_j"	DIMENSION DE LA REGLE $x \rightarrow y$
1	Fraîcheur	$min[q_1(x), q_1(y)]$	La fraîcheur de la règle $x \rightarrow y$ est estimée de façon pessimiste par le pire score de fraîcheur des 2 ensembles de données composant la règle.
2	Précision	$q_2(x) . q_2(y)$	La précision de la règle $x \rightarrow y$ est estimée par le produit des scores de précision des ensembles x et y de la règle.
3	Complétude	$q_3(x) + q_3(y) - q_3(x) . q_3(y)$	La complétude de la règle $x \rightarrow y$ est estimée par la probabilité qu'un des deux ensembles de la règle soit complet.
4	Cohérence	$max[q_4(x), q_4(y)]$	La cohérence de la règle $x \rightarrow y$ est estimée de façon optimiste par le meilleur score de cohérence des 2 ensembles de données composant la règle.

TAB. 1 – *Différentes fonctions de fusion pour combiner les scores d'une dimension qualité*

Nous considérons que choisir et utiliser (ou non) une règle d'association est une décision qui désigne la règle comme étant légitimement intéressante (D1), potentiellement intéressante (D2), ou inintéressante (D3) à la fois sur la base de « bonnes » mesures d'intérêt mais, également, en connaissant la qualité effective des données composant les parties gauche et droite de la règle. Cette décision a nécessairement un coût lié à l'incertitude et à la méconnaissance du jeu de données et du processus de recueil qui ne garantit généralement pas la bonne qualité des données. Pour formalisme, nous employons $P_{CE}(x)$ la probabilité que le jeu de données X soit "de mauvaise qualité" sur une ou plusieurs dimensions de qualité et $P_{CC}(x)$ la probabilité que X soit "de qualité correcte" dans une gamme de valeurs acceptables sur chaque dimension de qualité. $P_{AE}(x)$ représente la probabilité que X soit détecté correct alors qu'il est effectivement incorrect, $P_{AC}(x)$ représente la probabilité qu'il soit détecté incorrect alors qu'il est effectivement correct (voir Figure 1).

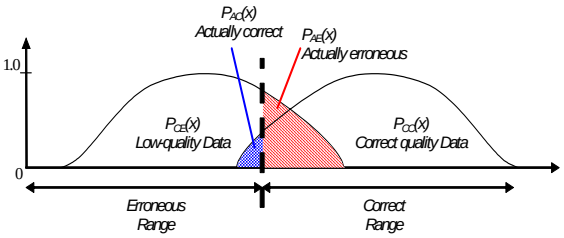

FIG.1 – *Probabilités de détection des données dont la qualité est correcte ou mauvaise*

Décision de sélection d'une règle	*Coût#*	*Statut de la qualité des données*	*Coût*
D_1	c_{10}	CC	\$0.00
	c_{11}	CE	\$1 000.00
	c_{12}	AE	\$0.00
	c_{13}	AC	\$0.00
D_2	c_{20}	CC	\$50.00
	c_{21}	CE	\$50.00
	c_{22}	AE	\$0.00
	c_{23}	AC	\$0.00
D_3	c_{30}	CC	\$500.00
	c_{31}	CE	\$0.00
	c_{32}	AE	\$0.00
	c_{33}	AC	\$0.00

TAB. 2 – *Coûts des décisions de sélection des règles selon la qualité des données*

Pour $\overline{q} \in Q$, la qualité moyenne des données $LHS \cup RHS$ de la règle R, on note $P(\overline{q} \in Q \mid CC)$ ou $f_{CC}(\overline{q})$ la probabilité conditionnelle que la qualité moyenne $\overline{q}$ corresponde à celle des jeux de données qui sont classifiés comme étant de qualité correct (CC). De la même façon, on note $P(q \in Q \mid CE)$ ou $f_{CE}(\overline{q})$ la probabilité conditionnelle que la qualité moyenne $\overline{q}$ corresponde à celle des jeux de données qui sont classifiés erronés ou comme étant de mauvaise qualité (CE). On note d la décision de choisir une règle légitimement intéressante (notée D_1), potentiellement intéressante (notée D_2), ou inintéressante (notée D_3) et s le statut de la qualité des jeux de données à partir desquelles a été calculée la règle. On note $P(d=D_i, s=j)$ et $P(d=D_i \mid s=j)$ respectivement les probabilités conjointe et conditionnelle que la décision D_i soit prise lorsque le statut de la qualité des données à l'origine du calcul de la règle R est j (*i.e.*, *CC, CE, AE, AC*). On note c_{ij} le coût de la décision D_i pour classifier la règle sur la base de la qualité des données j composant les parties gauche et droite de la règle. A partir des coûts présentés dans le Tableau 2, l'objectif est de minimiser le coût moyen $\overline{c}$ qui résulte d'une telle prise de décision tel que :

$$\begin{aligned}\overline{c} = {} & c_{10}.P(d=D_1, s=CC) + c_{20}.P(d=D_2, s=CC) + c_{30}.P(d=D_3, s=CC) \\ & + c_{11}.P(d=D_1, s=CE) + c_{21}.P(d=D_2, s=CE) + c_{31}.P(d=D_3, s=CE) \\ & + c_{12}.P(d=D_1, s=AE) + c_{22}.P(d=D_2, s=AE) + c_{32}.P(d=D_3, s=AE) \\ & + c_{13}.P(d=D_1, s=AC) + c_{23}.P(d=D_2, s=AC) + c_{33}.P(d=D_3, s=AC) \end{aligned} \quad (3)$$

A partir du théorème de Bayes, tel que:

$$P(d=D_i, s=j) = P(d=D_i \mid s=j).P(s=j) \text{ où } i=1,2,3 \text{ et } j=CC,CE,AE,AC \quad (4)$$

on suppose que $\overline{q}$ est la qualité moyenne des jeux de données composant la règle tirée aléatoirement dans l'espace de tous les vecteurs qualité possibles. La probabilité conditionnelle $P(d=D_i \mid s=j)$ est définie à partir de la fonction de densité de probabilité f_j du vecteur qualité de telle sorte que la variable aléatoire $\overline{q}$ prenne des valeurs dans l'intervalle $[q_{jMin}, q_{jMax}]$ correspondant aux seuils de qualité selon j telle que : $P(d=D_i \mid s=j) = \sum_{\overline{q} \in]q_{jMin}, q_{jMax}]} f_j(\overline{q})$. où $i=1,2,3$ et $j=CC,CE,AE,AC$ (5)

Par souci de place pour la suite de l'article, considérons le cas où il n'y a pas de problème de classification : $P(s=AC)$, $P(s=AE)$, f_{AC}, f_{AE} sont nuls et π^0_{AE} et π^0_{AC} éga-

lement. On note la probabilité a priori $P(s=CC)=\pi^0$, la probabilité a priori probabilité de $P(s=AC)=\pi^0_{AC}$, la probabilité a priori de $P(s=AE)=\pi^0_{AE}$ et la probabilité a priori de $P(s=CE) = 1-\pi^0$. Le coût moyen $\overline{c}$ dans l'éq. (3) basée sur les éq. (4) et (5) peut être réécrit tel que :

$$\begin{aligned}\overline{c} = &\sum_{\bar{q}\in Q_1}\left[f_{CC}.c_{10}.\pi^0 + f_{CE}.c_{11}.(1-\pi^0) + f_{AE}.c_{12}.\pi^0_{AE} + f_{AC}.c_{13}.\pi^0_{AC}\right]\\ &+\sum_{\bar{q}\in Q_2}\left[f_{CC}.c_{20}.\pi^0 + f_{CE}.c_{21}.(1-\pi^0) + f_{AE}.c_{22}.\pi^0_{AE} + f_{AC}.c_{23}.\pi^0_{AC}\right]\\ &+\sum_{\bar{q}\in Q_3}\left[f_{CC}.c_{30}.\pi^0 + f_{CE}.c_{31}.(1-\pi^0) + f_{AE}.c_{32}.\pi^0_{AE} + f_{AC}.c_{33}.\pi^0_{AC}\right]\end{aligned} \quad (6)$$

Minimiser ce coût moyen conduit à rechercher la décision optimale pour la classification des règles en trois catégories notées D^0_1 (pour les règles légitimement intéressantes), D^0_2 (pour les règles potentiellement intéressantes) et D^0_3 (pour les règles inintéressantes) selon les trois répartitions de qualité correspondantes Q_1, Q_2 et Q_3.

$$\begin{aligned}D^0_1 &= \left\{\bar{q} : \frac{f_{CE}}{f_{CC}} \le \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \le \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{20}-c_{10}}{c_{11}-c_{21}}\right\}\\ D^0_2 &= \left\{\bar{q} : \frac{f_{CE}}{f_{CC}} \ge \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{20}-c_{10}}{c_{11}-c_{21}} \text{ and, } \frac{f_{CE}}{f_{CC}} \le \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\right\}\\ D^0_3 &= \left\{\bar{q} : \frac{f_{CE}}{f_{CC}} \ge \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{10}}{c_{11}-c_{31}} \text{ and, } \frac{f_{CE}}{f_{CC}} \ge \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{20}}{c_{21}-c_{31}}\right\}\end{aligned} \quad (7)$$

Les inégalités (7) permettent de définir trois valeurs de seuil L, P et N (respectivement pour des règles légitimement, potentiellement et non intéressantes) dans l'espace de décision. L'éq. (8) définit concrètement ces régions où la décision est basée sur les coûts du choix d'une règle sans problème de classification:

$$L = \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{10}}{c_{11}-c_{31}},\ P = \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{20}-c_{10}}{c_{11}-c_{21}} \text{ et } N = \frac{\pi^0}{1-\pi^0}\cdot\frac{c_{30}-c_{20}}{c_{21}-c_{31}} \quad (8)$$

3 Expérimentations

Afin d'évaluer notre modèle décisionnel, nous avons utilisé les données recueillies par l'UCI dans le cadre de KDD-CUP-98[1] comprenant 191.779 enregistrements sur les individus contactés pour une campagne de dons aux vétérans américains en 1997. Chaque enregistrement est décrit par 479 variables et deux variables cibles indiquent les classes d'individus ayant répondu positivement ou non ("respond"/"no respond") avec la donation effectuée en dollars. Environ 5% des individus ont répondu positivement. Le challenge de KDD-Cup-98 était d'établir un modèle de prédiction des bénéfices, c'est-à-dire du montant total des donations auquel sont retranchés les coûts d'expédition $0.68 (voir (Wang *et al.* 2005) pour plus de détails). Puisque nous ignorions la qualité des données recueillies pendant cette campagne, nous avons généré des indicateurs synthétiques de la qualité de données selon différentes distributions de pollutions sur quatre dimensions de la qualité des données (fraîcheur, complétude, cohérence et crédibilité) pour chaque domaine de valeurs des variables considé-

[1] *http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html* for the dataset and *http://www.kdnuggets.com/meetings/kdd98/kdd-cup-98.html* for the results

rées. Puis, nous avons calculés les 10 meilleures règles d'association en suivant l'approche de Wang *et al.* (2005) pour prédire les individus répondant positivement à l'appel de dons et les bénéfices. Nous avons ensuite calculé la qualité des 10 meilleures règles découvertes (voir Tableau 3). En employant l'éq. (8) et le Tableau 2, nous avons calculé les valeurs des trois seuils de décision pour le choix des règles avec une probabilité a priori $\pi^0 = 0.200$ sans problème de classification et obtenons $L = 0.125$, $P = 0.013$ et $N = 2.25$.

#	*Régles d'associations*	*(Conf. ; Supp.)*	*Profit* (Wang et al., 2005)	*Qualité*					*Coût*	*Décision*
				Fresh.	*Accur.*	*Compl.*	*Consi.*	*Average*		
1	ETHC4=[2.5,4.5], ETH1=[22.84,29.76], HC6=[60.91,68.53]	(0.11; 13)	**$81.11**	0,21	0,38	0,79	0,53	0,48	**$53**	POTENTIALLY
2	RFA_14=f1d, ETH1=[29.76,36.69]	(0.17; 8)	**$61.73**	0,21	0,52	0,62	0,53	0,47	**$109.5**	NOT
3	HHD1=[24.33,28.91], EIC4=[33.72,37.36]	(0.12;12)	**$47.07**	0,17	0,35	0,90	0,15	0,39	**$113**	NOT
4	RFA_23=s2g, ETH13=[27.34,31.23]	(0.12;16)	**$40.82**	0,34	0,01	0,90	0,79	0,51	**$130**	NOT
5	EIC16=[11.25,13.12], CHIL2=[33,35.33], HC6=[45.69,53.30]	(0.16;11)	**$35.17**	0,03	0,53	0,77	0,71	0,51	**$34.7**	POTENTIALLY
6	RHP2=[36.72,40.45], AGE904=[42.2,44.9]	(0.16;7)	**$28.71**	0,50	0,15	0,44	0,73	0,46	**$109**	NOT
7	HVP5=[56.07,63.23], ETH13=[31.23,35.61], RAMNT_22=[7.90,10.36]	(0.14;10)	**$24.32**	0,37	0,65	0,68	0,95	0,66	**$62.8**	POTENTIALLY
8	NUMCHLD=[2.5,3.25], HU3=[66.27,70.36]	(0.08;31)	**$19.32**	0,07	0,09	0,61	0,57	0,34	**$190**	NOT
9	RFA_11=f1g, DMA=[743,766.8], POP903=[4088.208,4391.917], WEALTH2=[6.428571,7.714286]	(0.25;8)	**$17.59**	0,24	0,08	0,72	0,95	0,50	**$49.6**	POTENTIALLY
10	HUPA1=[41.81+,], TPE11=[27,64,31.58]	(0.23;9)	**$9.46**	0,20	0,22	0,99	0,93	0,59	**$40.8**	POTENTIALLY

TAB. 3 – *Les 10 meilleures règles découvertes avec leur qualité, coût et décision associés*

Le Tableau 3 montre les 10 meilleures règles d'associations découvertes avec leur confiance, leur support (en nombre d'enregistrements), le bénéfice prédit, les scores par dimension de qualité, la qualité moyenne, le coût de sélection pour chaque règle d'association et enfin, la région décisionnelle de chaque règle. Il est intéressant de noter que le bénéfice prédit par règle peut être considérablement affecté par le coût de sélectionner une règle calculée à partir de données de mauvaise qualité : par exemple, la deuxième meilleure règle R2 dont le bénéfice prévu est $61.73 a un coût de $109.5 si elle est sélectionnée alors qu'elle est classifiée comme "inintéressante" à cause de cette mauvaise qualité de données. Dans ces expériences, notre but était également de démontrer que des variations de la qualité de données pouvaient avoir un impact considérable sur la validité des résultats issus de la fouille de règles et dans le cas du challenge KDD-Cup-98, invalider totalement les prédictions de bénéfices. Ainsi, la Figure 2(a) montre le comportement du coût de la décision dans le choix des règles quand la qualité moyenne des données (`InitQual`) se dégrade de -10%, -30%, à -50% ou s'améliore de +10%, +30% à +50% avec $\pi^0 = 0.200$ et en l'absence de problème de classification. Nous observons que la dégradation de la qualité des jeux de données composant les règles augmente le coût de la sélection de ces règles. L'amélioration de qualité de données se manifeste par une stabilisation du coût de la décision pour des règles légitimement intéressantes. Un autre résultat intéressant est montré dans la Figure 2(b) : parmi les 10 meilleures règles découvertes pour une qualité de données initiale (`InitQual`), seulement 5 règles (R1, R5, R7, R9 et R10) sont potentiellement intéressantes. Lorsque la qualité augmente de 30%, 3 règles deviennent légitimement intéressantes (R5, R7 et R9). Cette observation offre deux perspectives intéressantes pour l'exploitation des règles d'associations et pour la gestion de la qualité de données : la première pour guider l'élagage du nombre de règles sur la base à la fois des indicateurs de qualité de données et des coûts de décision dans le choix des règles ; la seconde pour établir des stratégies et des priorités dans l'amélioration de la qualité de données (par un nettoyage ciblé par exemple).

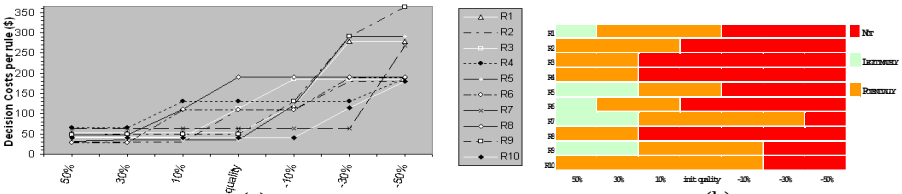

FIG. 2 – *Coût de la décision et statut des règles selon des variations de la qualité*

4 Conclusion

Dans cet article, nous proposons une méthode pour définir la qualité des règles d'associations en intégrant des mesures de la qualité de données à partir desquelles celles-ci sont découvertes. Ensuite, nous proposons un modèle décisionnel probabiliste basé sur le coût que peut engendrer le choix de règles d'association certes intéressantes d'après leur support et confiance mais basées sur des données de mauvaise qualité. Le modèle définit les trois seuils pour déterminer si les règles découvertes sont légitimement, potentiellement intéressantes, ou inintéressantes. Pour valider notre approche, nos expériences sur l'ensemble de données de la KDD-Cup-98 ont consisté à : *i)* générer des indicateurs synthétiques de qualité des données, *ii)* calculer les dix meilleures règles d'association (en terme de support/confiance) et calculer leur qualité moyenne à partir de la qualité de leurs données, *iii)* calculer le coût qu'entraîne la décision de choisir des règles illégitimement intéressantes, *iv)* examiner le coût et la décision sur le choix de ces règles quand la qualité des données varie. Nos expériences confirment notre hypothèse initiale : les mesures d'intérêt des règles d'associations découvertes ne sont pas autosuffisantes et la qualité d'une règle d'association dépend de la qualité des données à partir desquelles elle est calculée. La qualité de données inclut de diverses dimensions devant être considérées pour assurer et valider la qualité des connaissances extraites.

Références

Tan P-N., Kumar V., and Srivastava J., (2002). Selecting the Right Interestingness Measure for Association Patterns, *Proc. of Intl. KDD Conf.*, p. 32-41.

Wang K., Zhou S., Yang Q., and Yeung J.M.S., (2005). Mining Customer Value: from Association Rules to Direct Marketing, *Journal of Data Mining and Knowledge Discovery*.

Summary

In this paper we propose to exploit the measures describing the quality of data for defining the quality of association rules resulting from the rule mining process. We propose a probabilistic cost-based decisional model for the selection of legitimately, potentially interesting or not interesting rules when the quality of data they are computed from is correct, medium or low. The experiments on the KDD-CUP-98 dataset show that the Top 10 rules selected with good support and confidence measures are not interesting unless the quality of their data is correct or improved.