

***IPEE* : Indice Probabiliste d'Ecart à l'Equilibre pour l'évaluation de la qualité des règles**

Julien Blanchard, Fabrice Guillet
Henri Briand, Régis Gras

LINA – FRE 2729 CNRS
Polytech’Nantes
La Chantrerie – BP 50609
44306 – Nantes cedex 3 – France
julien.blanchard@polytech.univ-nantes.fr

Résumé. La mesure de la qualité des connaissances est une étape clef d’un processus de découverte de règles d’association. Dans cet article, nous présentons *IPEE*, un indice de qualité de règle qui a la particularité unique d’associer les deux caractéristiques suivantes : d’une part, il est fondé sur un modèle probabiliste, et d’autre part, il mesure un écart à l’équilibre (incertitude maximum de la conclusion sachant la prémisse vraie).

1 Introduction

Parmi les modèles de connaissances utilisés en Extraction de Connaissances dans les Données (ECD), les règles d’association (Agrawal et al., 1993) sont devenues un concept majeur qui a donné lieu à de nombreux travaux de recherche. Ces règles sont des tendances implicatives $a \rightarrow b$ où a et b sont des conjonctions d’items (variables booléennes de la forme *attribut = valeur*). Une telle règle signifie que la plupart des enregistrements qui vérifient a dans les données vérifient aussi b .

Une étape cruciale dans un processus de découverte de règles d’association est la validation des règles après leur extraction. En effet, de par leur nature non supervisée, les algorithmes de data mining peuvent produire des règles en très grande quantité et dont beaucoup sont sans intérêt. Pour aider le décideur (expert des données étudiées) à trouver des connaissances pertinentes parmi ces résultats, l’une des principales solutions consiste à évaluer et ordonner les règles par des mesures de qualité (Tan et al., 2004) (Guillet, 2004) (Lallich and Teytaud, 2004) (Lenca et al., 2004). Nous avons montré dans (Blanchard et al., 2004) qu’il existe deux aspects différents mais complémentaires de la qualité des règles : l’écart à l’indépendance et l’écart à ce que nous appelons l’équilibre (incertitude maximum de la conclusion sachant la prémisse vraie). Ainsi, les mesures de qualité se répartissent en deux groupes :

- les indices d’écart à l’indépendance, qui prennent une valeur fixe quand les variables a et b sont indépendantes ($n_{ab} = n_a n_b$) ;
- les indices d’écart à l’équilibre, qui prennent une valeur fixe quand les nombres d’exemples et de contre-exemples sont égaux ($n_{ab} = n_{a\bar{b}} = \frac{1}{2}n_a$).

Les mesures de qualité peuvent également être classées selon leur nature descriptive ou statistique (Lallich and Teytaud, 2004) :

	Indice d'écart à l'équilibre	Indice d'écart à l'indépendance
Indice descriptif	<ul style="list-style-type: none"> – confiance, – indice de Sebag et Schoenauer, – taux des exemples et contre-exemples, – indice de Ganascia, – moindre-contradiction, – indice d'inclusion... 	<ul style="list-style-type: none"> – coefficient de corrélation, – indice de Loevinger, – lift, – conviction, – TIC, – rapport de cote, – multiplicateur de cote...
Indice statistique		<ul style="list-style-type: none"> – intensité d'implication, – indice d'implication, – indice de vraisemblance du lien, – contribution orientée au χ^2, – rule-interest...

TAB. 1 – Classification des mesures objectives de qualité de règle

- Les indices descriptifs (ou fréquentiels) sont ceux qui ne varient pas avec la dilatation des effectifs (quand les effectifs sont augmentés dans la même proportion).
- Les indices statistiques sont ceux qui varient avec la dilatation des effectifs. Parmi eux, on trouve en particulier les mesures probabilistes, qui comparent la distribution observée des données à une distribution théorique, comme l'intensité d'implication (Gras, 1996) ou l'indice de vraisemblance du lien (Lerman, 1981).

A l'aide de ces deux critères, nous classifions les mesures de qualité de règle en quatre catégories. Comme le montre le tableau 1 (voir (Guillet, 2004) pour les références), il n'existe aucun indice statistique qui mesure l'écart à l'équilibre. Pourtant, les indices statistiques ont l'avantage de prendre en compte la taille des phénomènes étudiés. Statistiquement, une règle est en effet d'autant plus fiable qu'elle est évaluée sur un grand volume de données. Dans cet article, nous proposons donc un nouvel indice de qualité de règle qui mesure l'écart à l'équilibre tout en étant de nature statistique. L'indice est présenté dans la partie suivante, puis ses propriétés sont étudiées partie 3.

2 Mesure de la significativité statistique de l'écart à l'équilibre

Nous considérons un ensemble E de n objets décrits par des variables booléennes. Dans le vocabulaire des règles d'association, les objets sont des transactions enregistrées dans une base de données, les variables sont appelées des items, et les conjonctions de variables des itemsets. Etant donné un itemset a , nous notons A l'ensemble des transactions qui vérifient a , et n_a le cardinal de A . Une règle d'association est un couple (a, b) noté $a \rightarrow b$ où a et b sont deux itemsets qui ne possèdent pas d'item en commun. Les exemples de la règle sont les objets qui vérifient a et b , tandis que les contre-exemples sont les objets qui vérifient a mais pas b .

Etant donnée une règle $a \rightarrow b$, nous cherchons à mesurer la significativité statistique de l'écart à l'équilibre de la règle. La configuration d'équilibre étant définie par l'équirépartition dans A des exemples $A \cap B$ et des contre-exemples $A \cap \bar{B}$, l'hypothèse de référence est l'hypothèse H_0 d'équiprobabilité entre les exemples et les contre-exemples. Associons donc à l'ensemble A un ensemble aléatoire X de cardinal n_a tiré dans E sous

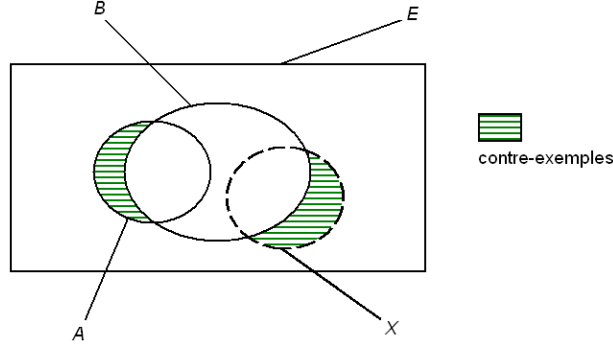


FIG. 1 – Tirage aléatoire d'un ensemble X sous hypothèse d'équiprobabilité entre les exemples et les contre-exemples

cette hypothèse : $P(X \cap B) = P(X \cap \bar{B})$ (voir figure 1). Le nombre de contre-exemples attendu sous H_0 est le cardinal de $X \cap \bar{B}$, noté $|X \cap \bar{B}|$. Il s'agit d'une variable aléatoire dont $n_{a\bar{b}}$ est une valeur observée. La règle $a \rightarrow b$ est d'autant meilleure que la probabilité que le hasard produise plus de contre-exemples que les données est grande.

Définition 1 L'indice probabiliste d'écart à l'équilibre (*IPEE*) d'une règle $a \rightarrow b$ est défini par :

$$IPEE(a \rightarrow b) = P(|X \cap \bar{B}| > n_{a\bar{b}} \mid H_0)$$

IPEE quantifie donc l'in vraisemblance de la petitesse du nombre de contre-exemples $n_{a\bar{b}}$ eu égard à l'hypothèse H_0 . En particulier, si $IPEE(a \rightarrow b)$ est proche de 1 alors il est invraisemblable que les caractères (a et b) et (a et \bar{b}) soient équiprobables. Ce nouvel indice peut être interprété comme le complément à 1 de la probabilité critique (*p-value*) d'un test d'hypothèse. Toutefois, à l'instar de l'intensité d'implication et de l'indice de vraisemblance du lien (où H_0 est l'hypothèse d'indépendance entre a et b), il ne s'agit pas ici de tester une hypothèse mais bien de l'utiliser comme référence pour évaluer et ordonner les règles.

Dans le cadre d'un tirage avec remise, $|X \cap \bar{B}|$ suit une loi binomiale de paramètres $\frac{1}{2}$ (autant de chances de tirer un exemple que de tirer un contre-exemple) et n_a . *IPEE* s'écrit donc :

$$IPEE(a \rightarrow b) = 1 - \frac{1}{2^{n_a}} \sum_{k=0}^{n_{a\bar{b}}} C_{n_a}^k$$

3 Propriétés de la mesure *IPEE*

Le tableau 2 présente les propriétés de *IPEE*. La mesure est également représentée en fonction du nombre de contre-exemples dans la figure 2. Nous pouvons voir que :

- *IPEE* réagit faiblement aux premiers contre-exemples (décroissance lente). Ce comportement est intuitivement satisfaisant puisqu'un faible nombre de contre-exemples ne saurait remettre en cause la règle.

Domaine de variation	$[0; 1]$
Valeur pour les règles logiques	$1 - \frac{1}{2n_a}$
Valeur pour les règles à l'équilibre	0.5
Variation avec $n_{a\bar{b}}$ pour n_a constant	\searrow
Variation avec n_a pour $n_{a\bar{b}}$ constant	\nearrow

TAB. 2 – Propriétés de la mesure *IPEE*

- Le rejet des règles s'accélère dans une zone d'incertitude autour de l'équilibre $n_{a\bar{b}} = \frac{n_a}{2}$ (décroissance rapide).

Comme le montrent les figures 3.(a) et (b), à proportion exemples/contre-exemples constante, les valeurs prises par *IPEE* sont d'autant plus extrêmes (proches de 0 ou 1) que n_a est grand. En effet, de par sa nature statistique, l'indice prend en compte la taille des phénomènes étudiés : plus n_a est grand, plus on peut avoir confiance dans le déséquilibre exemples/contre-exemples observé dans les données, et plus on peut confirmer la bonne ou la mauvaise qualité de l'écart à l'équilibre de la règle. Si la prise en compte de la taille des phénomènes étudiés fait la force des mesures de significativité statistique, ceci constitue aussi leur principale limite : elles sont peu discriminantes quand la taille des phénomènes est grande (de l'ordre de 10^4). En effet, au regard d'effectifs importants, même des écarts triviaux peuvent s'avérer statistiquement significatifs. *IPEE* ne déroge pas à la règle : quand n_a est grand, l'indice tend à évaluer que les règles sont soit très bonnes, soit très mauvaises.

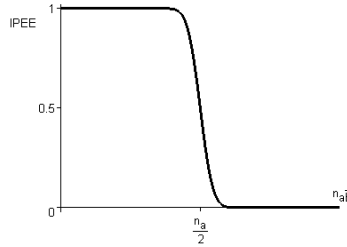
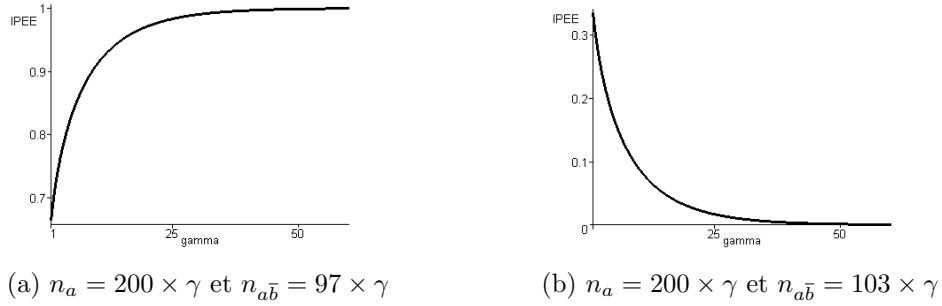


FIG. 2 – Variations de *IPEE* en fonction du nombre de contre-exemples $n_{a\bar{b}}$

4 Conclusion

Dans cet article, nous avons présenté *IPEE*, un nouvel indice de qualité de règle qui mesure l'écart à l'équilibre au regard d'un modèle probabiliste. De par sa nature statistique, cet indice a l'avantage de prendre en compte la taille des phénomènes étudiés, contrairement aux autres mesures d'écart à l'équilibre.

IPEE peut être vu comme l'analogue de l'intensité d'implication (Gras, 1996) pour l'écart à l'équilibre. Utilisées conjointement, ces deux mesures permettent une évaluation statistique complète des règles. La suite de ce travail de recherche consis-

FIG. 3 – Variations de *IPEE* avec la dilatation des effectifs

tera principalement à intégrer *IPEE* à notre plate-forme de validation de règles *ARVis* afin d'expérimenter le couple (*IPEE*, intensité d'implication) sur des données réelles.

Références

- Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on management of data*, pages 207–216. ACM Press.
- Blanchard, J., Guillet, F., Gras, R., and Briand, H. (2004). Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel tic. *Revue des Nouvelles Technologies de l'Information*, E-2 :287–298. Actes des journées EGC 2004.
- Gras, R. (1996). *L'implication statistique : nouvelle méthode exploratoire de données*. La Pensée Sauvage Editions.
- Guillet, F. (2004). Mesures de la qualité des connaissances en ecd. Tutoriel des journées Extraction et Gestion des Connaissances (EGC) 2004, www.isima.fr/~egc2004/Cours/Tutoriel-EGC2004.pdf.
- Lallich, S. and Teytaud, O. (2004). Evaluation et validation de l'intérêt des règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1 :193–218.
- Lenca, P., Meyer, P., Vaillant, B., Picouet, P., and Lallich, S. (2004). Evaluation et analyse multicritère des mesures de qualité des règles d'association. *Revue des Nouvelles Technologies de l'Information*, E-1 :219–246.
- Lerman, I. C. (1981). *Classification et analyse ordinale des données*. Dunod.
- Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4) :293–313.