# Indice Probabiliste Discriminant de Vraisemblance du Lien pour des Données Volumineuses

Israël-César Lerman*, Jérôme Azé**

*Irisa-Université de Rennes 1
Campus de Beaulieu
35042 Rennes Cédex
lerman@irisa.fr
**Laboratoire de Recherche en Informatique
Université Paris-Sud
91405 Orsay
aze@lri.fr

**Résumé.** On sait que l'indice probabiliste implicatif usuel de vraisemblance du lien évaluant de façon intrinsèque une règle d'association, n'est plus discriminant si le nombre d'observations augmente suffisamment. Le but de cet article est de montrer l'extension discriminante de cet indice probabiliste pour évaluer une règle d'association, mais, dans le contexte d'un ensemble de règles. Cette approche a été proposée de longue date et a été validée dans le cadre de la classification hiérarchique AVL (Analyse de la Vraisemblance des Liens) d'un ensemble d'attributs de types quelconques. Une analyse expérimentale qui consiste à faire croître la taille des données par l'adjonction de contre-exemples à tous les attributs, montre toute la pertinence de la démarche statistique. Cette dernière est, au préalable, justifiée conceptuellement.

**Mots Clés :** Règle d'association, indice probabiliste discriminant, validité.

## 1 Introduction

La donnée est un tableau d'incidence ou d'existence à double entrée de dimension $n * p$ croisant un ensemble $\mathcal{O} = \{o_i | 1 \leq i \leq n\}$ de $n$ objets avec un ensemble $\mathcal{A} = \{a^j | 1 \leq j \leq p\}$ de $p$ attributs booléens. On peut noter $\alpha_i^j$ la valeur « vrai » ou « faux » de l'attribut $a^j$ sur l'objet $o_i$ : $\alpha_i^j = a^j(o_i), 1 \leq i \leq n, 1 \leq j \leq p$. On code généralement 1 la valeur « vrai » et 0, la valeur « faux ». On peut supposer sans restreindre la généralité que la valeur « vrai » à un attribut est sémantiquement plus signifiante que la valeur « faux » à cet attribut. Cela, le plus souvent, se traduit statistiquement par le fait que le nombre d'objets où l'attribut est à « vrai » est inférieur au nombre d'objets où l'attribut est à « faux ». La représentation que nous adoptons d'un attribut booléen $a$ est son extension $\mathcal{O}(a)$ qui est le sous-ensemble des objets où $a$ est à « vrai ». Ainsi $\mathcal{A}$ se trouve représenté par un ensemble de parties de l'ensemble