ION: a pertinent new measure for mining information from many types of data

André Totohasina[(1)]
[(1)]Department of Mathematics and Computer science
High School for Teacher Training in Technique Education (ENSET)
University of Antsiranana, P.B. O – Antsiranana, 201-Madagascar
totohasina@yahoo.fr

Henri Ralambondrainy[(2)]
[(2)] Institute of Research on Mathematics and Computer Sciences and Applications (IREMIA)
Department of Mathematics and Computer science.
Faculty of Sciences and Technologies
University of La Réunion
15, Avenue René Cassin, P.B. 751, 97715, Saint-Denis, Messag Cedex 9, France
ralambon@univ-reunion.fr

Abstract

Since last decade, many methods with appropriate measures are proposed in knowledge discovery in databases. These measures aim at both improving the quality of mined association rules and reducing the problem of many nested rules. This paper presents a new statistical Implication Oriented Normalized measure, denoted ION. ION turns to be a unifying framework for several probabilistic measures of interestingness of association rules mined from diverse kind of dataset. It naturally leads to a pertinent algorithm for mining statistical implication, according to logical reasoning: one has the identity $ION(\neg b \rightarrow \neg a) \equiv ION(a \rightarrow b)$*, for any itemsets a and b. In addition, it takes into account both of positively or negatively oriented dependencies and of a deviation from equilibrium on large databases.*

Key words

Association rules (AR), Boolean or quantitative data, conditional expectation, interestingness, oriented correlation.

1. Motivation

The mining association rules problem was introduced by [2] before intensively studied [3], [8]. More theoretical works are also done [13]. Association rule mining is a data mining task that consists with discovering relationships among items or itemsets given a set of transactions. Briefly, AR mining problem is composed of two stages:

(1) Finding frequent itemsets, commonly respecting a given support threshold;
(2) Generating AR from these frequent itemsets, commonly with a given confidence threshold.

Motivated by usefulness of AR in many application fields such as diagnosis decision support, recommender systems, intrusion detection, etc., many algorithms concerned with improving performance in the treatment speedy are proposed in the literature (Apriori, AprioriTID and AprioriHybrid proposed by [3], CHARM, CLOSET, CLOSET+, etc.). However one obtains a huge amount of extracted AR affected of some drawbacks, for example, because of without taking into account of reference situations: independency, logical reasoning, deviation from equilibrium [7], etc. So, paradoxically, data mining itself provides a new knowledge management problem. In the other hand, not all rules with high support and confidence are interesting. In this paper, we consider the problem of finding objectively the most interesting rules from the set of all candidate association rules holding in a data. Often the most interesting ARs are those revealing unexpected information, or an additional predictive power. In this text, the traditional association rules (AR) of the type "a implies b" or "if a is true, then b will likely also true", denoted "$a \rightarrow b$", and called a positive association rule, are extended to those negative association of the three forms [4], "$a \rightarrow \neg b$", "$\neg a \rightarrow b$", and "$\neg a \rightarrow \neg b$" that are respectively called a right hand negative AR, left hand negative association rule and counter-opposite AR. In market-basket analysis, negative AR may help in task of identifying products that conflict each other or products that complement each other. In addition, for example about the two typical types of trading behaviours, as insider trading and market manipulation, that impair fair and efficient trading in securities stock markets, the market surveillance have to ensure a fair and trading environment for all participants through an alert system. Negative AR may assist in determining which alerts can be ignored. Suppose that each piece of evidence a, b, c, and d, can cause an alert of unfair trading e. Having the two rules "$a \rightarrow \neg e$" and "$c \rightarrow \neg e$", the team can make the decision of trading when a or c occurs: alert caused by a or c can be ignored. So the development of negative AR mining will allow companies to hunt more business chances. Notice that considering infrequent itemsets are there more useful than only taking account into frequent itemsets. Two key problems exist in negative AR mining [4]: (i) how to effectively identify interesting itemsets? (ii) How to effectively identify negative AR of interest? All traditional algorithms recalled above are only interested to positive ARs mining.

2. The probabilistic model. Related concepts and definitions

In this text, we consider a finite discrete probabilised space $(\Omega, S(\Omega), P)$, such that cardinality of Ω equals n, denoted $n = |\Omega|$, where $S(\Omega)$ denotes the set of all subsets of Ω, and P the intuitive probability such that for all event E in Ω, $P(E) = |E|/n$. Let Γ be the set of m boolean variables (also called attributs): $\Gamma = \{v_1, v_2, \ldots, v_m\}$; each boolean variable v_i is considered as a Bernoulli random variable defined on the sample space Ω such that $P(v_i = 1) = P(v_i^{-1}(1)) = 1/|v_i^{-1}(1)|$. Each non empty subset of Γ is called an *itemset*. For convenience, an itemset equally denotes a subset and an attribut or a variable. For typographic simplicity, for two itemsets u and v in $S(\Gamma)$, we shall write: $U = u^{-1}(1)$ the dual of u, $V = v^{-1}(1)$, $n_u = |U|$, $n_v = |V|$, $n_{uv} = |U \cap V|$, $\neg u$ = logical negation of u, $\overline{U} = \Omega - U$, $supp(u) = P(U)$ called the support of u (see next table 4). Notice that the couple (Ω, Γ) symbolises the matrix of Boolean data D with n lines and m columns. The notion of AR is defined as it was introduced for the first time by [2].

Definition 1 (support-confidence framework)*: An AR mined from the Boolean database D is a couple (u, v) of itemsets, denoted $u \rightarrow v$, such that $U \cap V = \varnothing$, $(Supp(u), Supp(v), Supp(u \cup v)) \geq ms$ and $P(V/U) \geq mc$, where ms and mc are two reels fixed in]0, 1[; we respectively call u, v, $Supp(u \cup v)$, the conditional probability $P(V/U) = Supp(u \cup v)/Supp(u)$, n_{uv} and $n_{u \rightarrow v}$, as the antecedent, the consequent, the support or cover, the confidence (conf), the number of examples and the number of counter-examples of the AR $u \rightarrow v$.*

An itemset whose support is greatest than the fixed threshold is called a frequent (also called large) itemset. How interpreting an AR? For example, in the market topic, suppose u and v are two frequent itemsets: when $supp(u \rightarrow v) = s$ and $conf(u \rightarrow v) = c$, one concludes that 100c% of transactions containing u also contain v, when 100s% of transactions contain u and v.

Definition 2: *A probabilistic quality measure is a real function μ defined on $Part(I^2)$ such that, for each AR $u \rightarrow v$, the real value $\mu(u \rightarrow v)$ depends exclusively on the four parameters n, Supp(u), Supp(v) and $Supp(u \cup v)$.*

Notice that the set equations $U = U \cap \overline{V} + U \cap V$ and $V = U \cap V + \overline{U} \cap V$ justify this sufficiency of four mentioned parameters in definition 2. In this work, we are interesting of finding additional probabilistic measures to avoid the above mentioned drawbacks provided by the exclusive sufficiency in Support-Confidence framework as criteria for mining AR. We shall try taking into account into coherence of dependency and surprise semantics.

3. Required principles for implicative measure

Inspired by the formal logical implication, where two propositions of the forms $(u \rightarrow v)$ *and* $(\neg v \rightarrow \neg u)$ are equivalent, that is to say they have equal logical values, we pose definitions as below.

Definition 3: *An AR quality measure μ is said a measure of implication (also called implicative measure), if for all* AR $u \rightarrow v$, *it verifies:*

$$\mu(\neg v \rightarrow \neg u) = \mu(u \rightarrow v).$$

Definition 4: *An AR quality measure μ is symmetric, if for each* AR $u \rightarrow v$, one has: $\mu(u \rightarrow v) = \mu(v \rightarrow u)$, *and perfectly symmetric, if $\mu(\neg u \rightarrow \neg v) = \mu(u \rightarrow v)$.*

An AR whose one of measures is implicative will be qualified an implicative association rule.

For example, the measure Support is symmetric, but not implicative. Confidence is a non symmetric and non implicative probabilistic measure. The basic principles (P_i) required for our approach are the first five criteria contained in following (P_1), (P_2) and (P_3).

(P_1) The three Piatetsky-Shapiro's principles [11]:

An interestingness measure of an AR $(u \rightarrow v)$ must be null in case of independency of antecedent and consequent, strictly increasing function of the number of examples when the three other parameters are fixed, and strictly decreasing function of the cardinality of antecedent dual or strictly decreasing function of the cardinality of consequent dual, when the three other parameters are fixed.

(P_2) A fourth Major & Mangano's principle [10]:

An interestingness measure of an AR $(u \rightarrow v)$ must be strictly increasing function of its cover, when the confidence is maintained constant greatest than fixed threshold.

(P_3) The fith Freitas'principle [9]:

A quality measure of an AR must be a non symmetric function.

Recent works show the necessity to verifying additional criteria for an AR quality measure, as recalled below:

(P_4) An AR quality measure must be strictly decreasing function and preferably concave depending on the number of counter-examples.

(P_5) For all logical rule, that is to say a rule without counter example, an AR quality measure must be constant.

(P_6) An AR quality measure must be easy for making threshold significant.

(P_7) An AR quality measure must be intelligible, providing rules with easy interpretation.

(P_8) Sensitivity to n: the quality measure should vary when data dilates.

(P_9) Due to [4]: A quality measure should make reference to deviation of uncertainty (also called deviation of equilibrium), that is to say in case of equality between number of examples and number of counter examples, the quality measure must be constant.

4. Statistical Implication Oriented Normalized (ION) quality measure of association rules

Let be $u \rightarrow v$ an AR. $U = u^{-1}(1)$, $V = v^{-1}(1)$ and $U \cap V$ are the corresponding events to respectively itemsets u, v

and u∪v. About the concept of conditional probability, one has the following intuitive states:

I_1. u and v are statistically independent, if P(V/U)=P(V).

I_2. u and v are positively (resp. negatively) dependent (also called in attraction situation (resp. in repulsion situation)), if P(V/U)>P(U) (resp. P(V/U)<P(V)):

in this case, one has $0< P(V/U) - P(V) \leq 1 - P(V)$ (resp. $- P(V) \leq P(V/U) - P(V) < 0$). Commonly, the inequality P(V/U)>P(V) (resp. P(V/U)<P(V)) is interpreted as v favouring u (resp. u disfavouring v). Notice that P(V/U)<P(V) is equivalent to $1 - P(V/U) > 1 - P(V)$, say $P(\overline{V}/U) > P(\overline{V})$. In the other terms, (u disfavours v) is equivalent to (u favours ¬v), say considering the right hand negative rule ($u \rightarrow \neg v$).

I_3. u and v are incompatible, if P(V/U) = 0.

I_4. The logical implication of u on v corresponds to the inclusion $U \subset V$ and to P(V/U) = 1.

By virtue of the continuity of P(V/U) as a function of $t = |U \cap V|$, the logical implication state is the above-limit of positive dependency (i.e. the mutual attraction) between u and v. So by duality, for taking account into negative dependency, in case of one of the two itemsets disfavours the other, then one obtains a negative rule, and for coherence, a quality measure of the initial positive rule should be a negative value, equal to - 1 of course in case of incompatibility. This heuristic provides the definition below.

Definition 5: *A probabilistic quality measure μ is called normalized and centered, if μ verifies the three Piatetsky-Shapiro (P1) principles, is non symmetric, and such that for any AR ($u \rightarrow v$) from a given context, one has: $\mu(u \rightarrow v) > 0$, if u favours v; $\mu(u \rightarrow v) < 0$, if u de favours v; $\mu(u \rightarrow v) = +1$, in case of logical implication; $\mu(u \rightarrow v) = -1$, in case of incompatibility.*

Since $n_{u\neg v} = n - n_{uv}$, one has the

Proposition 1: *All normalized quality measure of AR is a strictly decreasing function of the number of counter examples.*

The probabilistic normalized and centered measure μ_n, deduced from probabilistic quality measure μ, is called the normalized quality measure of μ. For example, it is easy to verify that Conviction is an implicative measure not normalized, because of taking infinite value in case of logical implication.

Theorem and definition 6:

The probabilistic quality measure defined as a conditional probability increment ratio such that for any AR ($u \rightarrow v$), one has:

$$ION(u \rightarrow v) = \begin{cases} \dfrac{P(V/U) - P(V)}{1 - P(V)}, & \text{if u favours v;} \\ \dfrac{P(V/U) - P(V)}{P(V)}, & \text{else,} \end{cases}$$

with $\{u, v\} \neq \{\varnothing, \Gamma\}$, is a normalized, centered, non symmetric and implicative quality measure (Cf. §6, §7). One calls it simply *the statistical Implication Oriented Normalized (ION) quality measure.*

Proof. Let us remark the equivalence between the two propositions " u disfavours v" and "u favours ¬v". So it is sufficient to proof the first half of the definition 6. Let be an AR ($u \rightarrow v$) such that u favours v. One has:

$$\begin{aligned} ION(\neg v \rightarrow \neg u) &= [P(\overline{U}/\overline{V}) - P(\overline{U})]/(1 - P(\overline{U})] \\ &= [1 - P(U/\overline{V}) - 1 + P(U)]/P(U) \\ &= [P(U) - P(U \cap \overline{V})/(1 - P(V))]/P(U) \\ &= [-P(U)P(V) + P(U \cap V)]/[P(U)(1 - P(V)] \\ &= [P(V/U) - P(V)]/(1 - P(V)] = ION(u \rightarrow v). \end{aligned}$$

Then ION is effectively an implication quality measure. In the event of a logical implication $u \rightarrow v$, it is easy to obtain that $ION(u \rightarrow v) = 1$. In case of independency, $ION(u \rightarrow v) = 0$. *In* case of incompatibility, $ION(u \rightarrow v) = -1$. ION is non symmetric, because equality $ION(v \rightarrow u) = ION(u \rightarrow v)$ *holds, if and only if P(V/U)=P(V/U) and P(U)=P(V).*

Important remark. It is interesting to notice that the explicit expression $ION(u \rightarrow v) = (n.n_{uv} - n_u n_v)/(n_u(n - n_v))$ shows both the increasing function of number of examples, decreasing function of number of counter examples, concave decreasing function of n_v, respectively in maintaining the three other parameters constant. The implicative property of ION allows that if an AR $u \rightarrow v$ is valid, then also $(\neg v \rightarrow \neg u)$. That is comparable of logical reasoning. Last, from above statement I_2 one has: $-1 \leq ION(u \rightarrow v) \leq 1$, and :

In case of positive (resp. negative) dependency, the bigger the $ION(u \rightarrow v)$ (resp. $- ION(u \rightarrow v)$) value, the higher the positive (resp. negative) dependence which tend to logical rule (resp. incompatibility).

It is easy to observe that the traditional probabilistic quality measure *Confidence* is not implicative. So for aiming at mining AR with degree of logical implication background, it is necessary to combine it with ION. Notice that ION has no probability background. *Confidence* and ION play complementary roles in mining AR task. *Is this combination optimal? That is an open problem.* Next we'll see relation between ION and some probabilistic quality measures of AR through normalization action.

5. Probabilistic measure normalization process

Let be μ a probabilistic quality measure of AR, μ_n its normalized. Let be $u \rightarrow v$ an AR. Let us denote:

- x_f and y_f: the coefficients corresponding to case "u favours v", depending on probabilities P(U) and P(V);
- x_d and y_d: the coefficients corresponding to case "u disfavours v", depending on P(U) and P(V).

Taking into account the continuity of the evolution in the two zones of attraction and repulsion (Figure 1), one should obtain:

$$\mu_{n(u \rightarrow v)} = \begin{cases} x_f.\mu(u \rightarrow v) + y_f, & \text{if u favours v;} \\ x_d.\mu(u \rightarrow v) + y_d, & \text{if u disfavours v;} \end{cases} \quad (S0)$$

Under the terms of continuity, these four coefficients are determined by passages at the limiting points. From where:

(i)Logical implication and independence give:

$$\begin{cases} x_f.\mu(u \to v)_{imp} + y_f = +1 \\ x_d.\mu(u \to v)_{ind} + y_d = 0 \end{cases} \quad (S1)$$

(ii) Independence and incompatibility imply:

$$\begin{cases} x_d.\mu(u \to v)_d + y_d = 0 \\ x_d.\mu(u \to v)_{ind} + y_d = -1 \end{cases} \quad (S2)$$

These two linear systems give:

$$\begin{cases} x_f = \dfrac{1}{\mu(u \to v)_{imp} - \mu(u \to v)_{ind}} \\ y_f = -\dfrac{\mu(u \to v)_{ind}}{\mu(u \to v)_{imp} - \mu(u \to v)_{ind}} \end{cases} \quad (s3)$$

and

$$\begin{cases} x_d = \dfrac{1}{\mu(u \to v)_{ind} - \mu(u \to v)_{inc}} \\ y_d = -\dfrac{\mu(u \to v)_{ind}}{\mu(u \to v)_{inc} - \mu(u \to v)_{ind}} \end{cases} \quad (S4)$$

This shows possibility of normalization-centering of a probabilistic quality measure of the rules. Reciprocally, from relation (S0) results the expression of initial measure μ according to its normalized centered μ_n:

$$\mu(u \to v) = \begin{cases} \dfrac{\mu_n(u \to v) - y_f}{x_f}, & \text{if u favours v;} \\ \dfrac{\mu_n(u \to v) - y_d}{x_d}, & \text{if u disfavours v.} \end{cases} \quad (s5)$$

By definition, its normalized centered evolves/moves according to the shéma (Figure 1) below.

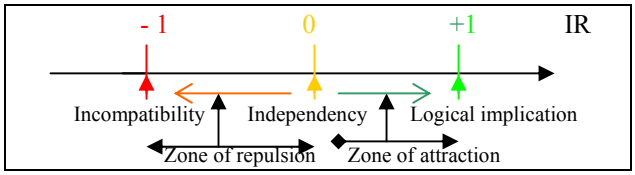

Figure 1 :Quality measure normalization process

Examples:

a) **Lovinger's measure**: Lov(u→v)=(P(V/U)-P(V))/(1-P(V)). Let be an association rule u →v. One has: if u favours v, then Lov(u→v)>0; in case of logical implication, one obtains Lov(u→v)=+1; in case of independence, Lov(u →v)=0; if u disfavours v, then Lov(u →v)<0; For the limit case of incompatibility, let us seek a function a checking $Lov_n(u \to v) = a.Lov(u \to v) = -1$:

- a P(V)/(1-P(V)) = - 1 implies a = (1-P(V))/P(V).

From there, results

$$Lov_n(u \to v) = \begin{cases} \dfrac{P(V/U) - P(V)}{1 - P(V)}, & \text{if u favours v;} \\ \dfrac{P(V/U) - P(V)}{P(V)}, & \text{else,} \end{cases}$$

Thus, Lov_n=ION. Conversely,

$$Lov(u \to v) = \begin{cases} ION(u \to v), & \text{if u favours v;} \\ \dfrac{P(V)}{1 - P(V)} ION(u \to v), & \text{else,} \end{cases}$$

b) By a similar reasoning, one obtains an invertible functional relation between ION and each following quality measures: confidence, conviction, Pearson's Phi-correlation coefficient, Piatetsky-Shapiro's measure, Surprise; and their normalized measures are equal to ION. This last proposition does not always hold for any probabilistic quality measure. The normalizability condition is precised below.

Proposition : *A probabilistic quality measure μ is normalizable, if and only if μ doesn't become infinite at one among the three references situations, that are incompatibility, independence or logical implication situations.*

However, if possible, these invertible relations with ION would provide an opportunity to compare many probabilistic quality measures via ION. From where results unifying property of ION for such AR quality measures.

6.ION sensitivity of references situations: independency, deviation from equilibrium and surprise. Significativity

Recall the objective interestingness measures may be divided into two groups: The measures taking account into deviation from independence, which have a fixed value when the two itemsets are independent, and those taking account into deviation from equilibrium (i.e. maximum uncertainty of the consequent given antecedent), which have a fixed value when number of examples and number of counter examples are equal.

Independency. For any possible AR u→v, one knows that the quantity P(V/U)-P(V))/P(V) = (P(V∩U)-P(U)P(V))/P(U)P(V), denoted δi(u,v), measures the deviation of independency ratio of the two itemets u and v, in case of u favours v. However in this case, ION(u→v)= δi(u,v). P(V)/(1-P(V)). Thus ION(u→v)> δi(u,v), if and only if P(V)/(1-P(V))>1, that is to say P(V)>1/2. Interpretation: *ION is an indicator of reduction of the uncertainty of v knowing the realization of u*, in case of P(V)>1/2.

Deviation from Equilibrium. Assume an AR u→v such that u favours v in equilibrium situation. Then $ION(u \to v)_{equilibrium} = 1/2 - n_v/(2(n - n_v)) = 1/2 - o(1/n)$ which tends to 1/2 when n becomes sufficiently big. Interpretation: *ION takes account into equilibrium deviation in large databases.*

Notice that $Conf(u \to v)_{equilibrium} = P(V/U) = 1/2$ which corresponds to effectively a maximum uncertainty. Inequation $ION(u \to v) > ION(u \to v)_{equilibrium}$ if and only if $Conf(u \to v) > Conf(u \to v)_{equilibrium} = 1/2$.

Surprise. Recall surprise [5]brought by an AR u→v is the quantity defined by

$Surp(u \to v) = (P(UV) - P(U \cap \overline{V}))/P(V)$.

Thus $Surp(u\rightarrow v)_{equilibrium}= 0$; and $Surp(u\rightarrow v)= 0$ if and only if P(V/U) =1/2: No surprising rule corresponds to maximum uncertainty of consequent given antecedent.

Interpretation: *Again, one would thus rather be brought to preserve only the rules whose confidence exceeds 1/2. The presence of P(V) to the denominator makes granting more credit to the rules whose support of consequent is relatively weak.*

One verifies that: $Surp_n$ = ION, $Sup((u\rightarrow v)= \begin{cases} \frac{P(V)ION(u\rightarrow v)}{2P(U)(1-P(V)} + 1 - \frac{1}{2(1-P(V))}, \text{ if u favours v,} \\ \frac{ION(u\rightarrow v)}{2P(U)} + \frac{1-2P(V)}{2P(V)}, \text{ else} \end{cases}$

and if $ION(u\rightarrow v)>0$ and $P(V)<1/2$, then $Surp(u\rightarrow v)>0$. *Thus ION takes into account of surprise background.*

Significativity. Remark that contingency table is implicitly referenced in defining ION. It is easy to show that ION, famous Khi-square and Pearson's Φ^2 statistics are linked by:

$$ION(u\rightarrow v) = \frac{n.n_{uv} - n_u n_v}{n_u(n-n_v)} = \sqrt{\frac{n_v}{n_u}\frac{n-n_u}{n-n_v}}\Phi(u\rightarrow v)$$
$$= \pm\sqrt{\frac{1}{n}\frac{n_v}{n_u}\frac{n-n_u}{n-n_v}}\chi^2$$

Interpretation: *Therefore, the thresholds of significativity of ION can be obtained easily starting from those of Khi-square. ION satisfies principle (P_6).* Notice that ION checks all above evoked principles P_7, P_8 and P_9.

Negative rules versus positive rules. For any possible AR u→v, one has :

$ION(u\rightarrow\neg v) = - ION(u\rightarrow v)$, and for any $\alpha\in]0, 1[$, $\alpha<ION(u\rightarrow v)<1 \Leftrightarrow - 1<ION(u\rightarrow\neg v)< - \alpha$.

Examples justifying complementarity of different criteria, and why ION is pertinent.

(i) Resumption of contingency table in S. Brin & al [8]:

Table1	v	¬v	
u	**20**	5	**25**
¬u	70	5	**75**
	90	**10**	100

$Conf(u\rightarrow v)=80\%>conf(u\rightarrow v)=22\%$,
$Conf(\neg u\rightarrow v)=93\%>conf(v\rightarrow\neg u)=78\%$
$ION(u\rightarrow v) < 0$: (u→v) is not valid.
$ION(u\rightarrow\neg v)=1/9 = 0, 11$: u→¬v is out of interest
$ION(\neg u\rightarrow v)=1/3 = 0, 33$: ¬u→v is of interest

(ii) Case of large databases:

Table 2	v	¬v	
u	**5 000 000**	500 000	**5500 000**
¬u	4 000 000	500 000	**4 500000**
	9000 000	**1000 000**	10000000

$Conf(u\rightarrow v)=90,90\%$, $Supp(u\rightarrow v)=90\%$, $Surp(u\rightarrow v)=0,50$, but $ION(u\rightarrow v) = 0, 09$ is weak: (u→v) is not of interest.

Table 3	v	¬v	
u	**5 000 000**	500 000	**5500 000**
¬u	4 000 000	500 000	**4 500000**
	9000 000	**1000 000**	10000000

$Conf(u\rightarrow v)=59\%$, $ION(u\rightarrow v) =0,69\%$,and $\Phi(u\rightarrow v)=67,7\%$, but $Supp(u\rightarrow v)=06, 8\%$ is very weak: (u→v) is not valid.

Now let us analyze the binary context defined below in Table 4: consider u={a, c} and v={b, e}.

Table 4	a	b	c	d	e
e_1	1	0	1	1	0
e_2	0	1	1	0	1
e_3	1	1	1	0	1
e_4	0	1	0	0	1
e_5	1	1	1	0	1

U={e_1, e_3, e_5}, V={e_2, e_3, e_4, e_5}, *Conf(u→v)=0.67,Supp(u→v)*=0.4,

P(V/U)=0.7<P(V)=0.8: The rule *u→v would be valid according the support-confidence pair, but not according the to ION, because it is shown that in fact u disfavours v. So it is advisable to consider the negative candidate AR (* u→¬v).

These four typical examples above illustrate some weakness in the exclusive support-confidence framework and the need for considering ION.

7. Related work

To our knowledge, few work are interested in the negative rules in addition to those positive.

First, we discuss the approach proposed by Wu et al.[12]. These authors define a new concept of *confidence* which is equal to the conditional probability increasing ratio, i.e. which, as by chance, coincides with our current measure suggested ION: it is called Confidence conditional probability increment ratio function, denoted *Confidence CPIR*. But ION does not contain no probabilistic background. In addition their algorithm automatically output negative AR of the type (¬u→¬v) without consideration of its logical equivalent u→v, but it is actually shown that ION or CPIR is implicative. In this approach, incompatibility is confused wrongly with the negative dependence.

Second, by strongly criticizing the approach of [12], work of Antonie and Zaïne[4] proposes an approach which combines the traditional model support-confidence with the coefficient of correlation in order to primarily extract stronger positive and confined negative AR, but coefficient correlation, which is intrinsically symmetric, is not implicative.

8. Mining implicative AR algorithm

The properties of ION studied above allow us to lead to the following definition of AR.

Definition 7: *Let $\Gamma = \{v_1, v_2, ..., v_m\}$ be the set of items in database D, S(Γ) be the set of all subsets of Γ, $w=u\cup v$ be an itemset such that $u\cap v =\varnothing$, $supp(u)\times supp(v)\neq 0$, and ms is a fixed real in]0, 1[and $\alpha\in]85,$*

1[given threshold by the user. Then, respecting above adopted notations:

(1) If supp(w)≥ms, supp(u)≥ms, supp(v)≥ms, conf(u→v)>min(1/2, conf(v→u)) and ION(u→v) α-significant then (u→v) is a positive rule of interest.

(2) If supp(u∪¬v)≥ms, supp(u)≥ms, supp(v) ≥ms, conf(u→¬v)>min(1/2, conf(¬v→u)) and ION(u→¬v) α-significant then (u→¬v) is a right hand negative AR of interest.

(3) If supp(u∪¬v)≥ms, supp(u)≥ms, supp(v) ≥ms, conf(¬v→u)>min(1/2, conf(u→¬v)) and ION(¬v→u) α-significant then ((¬v→u) is a left hand negative AR of interest.

Thus follows the suggested algorithm of mining implicative AR.

Stage 1: Extract frequent itemsets according welknown Apriori algorithm or its reviewed version; and for all disjoined frequent itemset u and v, by posing w=u+v, the second key sequence is described as follows.

Stage 2:

Case of w frequent itemset, **then**
if conf(u→v)>conf(v→u), **then**
E1: **if** supp(V)<1/2 **then**
if ION(u→v) significant, **then**
if conf(u→v)>1/2, **then** output the AR u→v,
else no AR,
else no AR,
else no AR,
else if conf(u→v)< conf(v→u), **then**
exchange u and v and go to **E1**,
else if conf(u→v)=conf(u→v)>1/2, **then**
if ION(u→v) and ION(v→u) both significant, **then** output equivalence u↔v,
else no rule,
Case of w infrequent, **then**
E2: **if** (u∪¬v) frequent, **then**
if ION(u→¬v) significant and Conf(u→¬v)>1/2, **then** output AR u→¬v ,
else if (¬u∪v) frequent, **then**
if ION(¬u→v) significant and Conf(¬u→v)>1/2, **then** output AR ¬u→v.

9. Concluding remarks

This study shows the existence of a probabilistic quality measure of ARs, denoted ION, which turns to play a central role compared to the usual probabilistic indices to assess the quality of association rules with interpretable dependence directed in statistical term of implication. This quality measure has a role comparable with that of the normal law centered and reduced within the Gaussian laws in the fields of the random variables. ION would be advantageously used by expert user who tolerates counter examples; and it would like belonging to the criteria in the case of approaches hybrid in the mining association rules task. ION can be helpfully used for pruning pertinently an AR basis.

10. References

[1] R.Agrawal, Heikki Mannila, Ramakrishman Srikant, Hannu Toivonen and A. Inkeri Verkamo (1996). *Fast Discovery of association rules*. *In* Advances in knowledge discovery and Data Mining, Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padraic Smyth, and Ramasamy Ulthurusamy Editors, AAAI Press / The MIT Press, California, pp. 308-328.

[2] R. Agrawal, T. Imielinski & A. Swami (1993). *Mining association rules betwen sets of items in large databases.* In P. Buneman and S. Jajodia, editors, Proc. Of ACM SIGMOD International Conference on Management of Data, volume 22, pp. 207-216, Wachington, 1993. ACM press.

[3] R. Agrawal and R. srikant (1994). *Fast algorithm for mining association rules.* In Proc. Of the 20th VLDB Conférence, 487-499.

[4]Antonie M.-L., Zaïane O.-R.(2004): Mining positive and negative Association Rules: an approach for confined rules. Technical Report TR04-07, Dept of Computing sciences, University of Alberta. http://ftp.cs.ualberta.ca/pub/TechReports/2004/TR04-07/TR04-07.ps. Avalable online

[5]Azé Jérome (2003). *Une nouvelle mesure de qualité pour l'extraction de pépites de connaissances, Extrcation des connaissances et apprentissage.* RSTI série RIA-ECA. Volume 17- n°1-2-3-/2003, Extraction et gestion des connaissances EGC 2003, pp.171-182.

[6]J. Blanchard, F. Guillet, H. Briand, and R. Gras, *Assessing rule interestingness with a probabilistic measure of deviation from equilibrium*, in Proc. ASMADA, ENST Bretagne, France, Mai 05, 191-200.

[7]Brin S., Motwani R. & Ulman J.D., & Tsur S.(1997). *Dynamic itemset counting and implications rules for market basket data.* Proc. Of the 1997 ACM SIGMOD conf, mai 1997b, 255-264.

[9]Freitas A.A. *On rule of interestingness measure.* Knowledge –Based-Systems n°12, 1999, 309-315. [

[10]J.A. Major and J.J. Mangano (1993), *Selecting among rules induced from a heuricane database*. In KDD'93, Workshop papers, pages 28-41, Menlo Park, Califoria.

[11]Piatetsky-shapiro G. *Knowledge discovery in Real Data Bases*. A report on the IJCAI-89 W.shop, AI Magazine, 11(5), 91, 68-70.

[12]Wu, X., Zhang C., Zhang S.(2004): Mining both positive and negative association rules. In ACM Transaction on Information Systems, Vol. 22, No 3, July 2004, p. 381-405.

[13] M.J. Zaki & M.Ohighara (1998). *Theoritical foundations of association rules.* In 3rd SIGMOD'98 workshop on Research Issues In data Mining and Knowledge Discovery (DMKD), pp. 1-8, 1998.