Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2007)

An Analytical Evaluation of Objective Measures Behavior for Generalized Association Rules

Veronica Oliveira de Carvalho
Professor of Centro Universitário de Araraquara
Araraquara, SP, Brazil
Student of São Paulo University
Computer and Mathematics Science Institute
São Carlos, SP, Brazil, 13560-970, Postal Code 668
Email: {veronica}@icmc.usp.br

Solange Oliveira Rezende, Mário de Castro
São Paulo University
Computer and Mathematics Science Institute
São Carlos, SP, Brazil, 13560-970, Postal Code 668
Email: {solange,mcastro}@icmc.usp.br

***Abstract*— The association rule mining task identifies all the intrinsic associations among the items contained in data and leads to only specialized knowledge. To overcome this problem the generalized association rules appeared. This type of rule associates not only the items contained in data, but also some items encoded into a given taxonomy. Therefore, the techniques used to obtain generalized association rules are very useful since they provide a more general view of the domain. However, a problem found when using these techniques is how to identify the most useful rules to avoid overload the user with a huge amount of patterns. Nowadays, the researches use objective evaluation measures to evaluate and select the most interesting knowledge to the user. Despite the fact these measures have been studied by many researches to evaluate many types of rules (for example, classification and traditional association rules), it is important to study these measures in the context of generalized rules. Thus, this paper presents an analytical evaluation to understand the behavior of some objective measures when applied in a set of generalized rules. Many relations were obtained to express the behavior of these measures, what represents a meaningful contribution to the post-processing data mining area.**

I. Introduction

One of the tasks in data mining is the *association rule mining*, which was introduced in [1] as follows. Consider D a database composed by a set of items $I = \{i_1, ..., i_m\}$ and by a set of transactions $T = \{t_1, ..., t_n\}$, where each transaction $t_i \in T$ is composed by a set of items (*itemset*), where $t_i \subseteq I$. A transaction $t_i \in T$ supports an itemset $A \subset I$ if $A \subset T$ holds. The fraction of transactions T supporting an itemset A with respect to database D is called the support of A, and is defined as $sup(A) = \frac{|\{T \in D; A \subset T\}|}{|D|}$ [2]. The support can be interpreted as the probability $P(E_A)$, where E_A is an event representing "*itemset A occurs*." An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, $A \cap B = \varnothing$. Two measures always appear with an association rule: support and confidence. The support of a rule $A \Rightarrow B$ is defined as $sup(A \Rightarrow B) = sup(A \cup B)$ [2], and can be interpreted as the probability $P(E_A \cap E_B)$. The confidence of a rule, that can be understood as the conditional probability $P(E_B|E_A)$, is defined as $conf(A \Rightarrow B) = \frac{|\{T \in D; A \cup B \subset T\}|}{|\{T \in D; A \subset T\}|}$ [2]. The problem of discovering all association rules is decomposed into two problems [1]: (a) find all large itemsets; (b) use these large itemsets to generate the rules. A large itemset is a set of items that has a support value no less than a user-specified minimum support (*minsup*). According to the rule generation step (item b), a rule will be generated, based on a given itemset, if the rule has a minimum user-specified confidence (*minconf*).

The use of a background knowledge in the data mining process allows the discovery of a more abstract, compact and, sometimes, interesting knowledge. An example of background knowledge can be a concept hierarchy, that is, a structure in which high level abstraction concepts (generalizations of low level concepts) are hierarchically organized by a domain expert or by an automatic process. An example of a simple concept hierarchy is taxonomy. Since the association rule mining technique generates all possible rules considering only the items contained in the data set, which leads to specialized knowledge, the *generalized association rules*, which are rules composed by items contained in any level of a given taxonomy, were introduced by [3]. Taxonomies reflect arbitrary individual or collective views according to which the set of items is hierarchically organized [4]. In the context of generalized association rules, a set of taxonomies is represented by a directed acyclic graph $\mathcal{T}$ on the transactions items, where an edge in $\mathcal{T}$ represents an *is-a* relationship.

Considering the same assumptions made in the traditional association rules, a generalized association rule is an implication of the form $A \Rightarrow B$, where $A \subset I$, $B \subset I$, $A \cap B = \varnothing$ and *no item in B is an ancestor of any item in A* [3]. In this case, the support and confidence measures are also used and there are some important relations that holds for those measures: (a) $sup(A \Rightarrow \widehat{B}) \geqslant sup(A \Rightarrow B)$; (b) $sup(\widehat{A} \Rightarrow B) \geqslant sup(A \Rightarrow B)$; (c) $sup(\widehat{A} \Rightarrow \widehat{B}) \geqslant sup(A \Rightarrow B)$; (d) $conf(A \Rightarrow \widehat{B}) \geqslant conf(A \Rightarrow B)$ [4], where $\widehat{A}$ indicates that $\widehat{A}$ is an ancestor of A and $\widehat{B}$ indicates that $\widehat{B}$ is an ancestor of B.

Taxonomies can be used in the different steps of the data mining process. Nowadays, many researches propose the obtaining of generalized association rules in the mining step ([3], [5], [6], [7], [8], [9], [10]) and in the pre-processing step ([11]). There are also some researches that apply taxonomies in the post-processing step ([12], [13], [14], [15]). In some

of those researches a generalized rule can be associated with a specialized rule. Let $A \Rightarrow B$ be a rule. The notation $\widehat{A \Rightarrow B}$ (resp., $\widetilde{A \Rightarrow B}$) represents new rules that derive from $A \Rightarrow B$ by replacing one or several items by its ancestors (resp., descendants) in $\mathcal{T}$. The new rules are said to be *generalizations* (resp., *specializations*) of $A \Rightarrow B$ [4]. Note that $\widehat{A \Rightarrow B}$ (resp., $\widetilde{A \Rightarrow B}$) stands for either $A \Rightarrow \widehat{B}$, $\widehat{A} \Rightarrow \widehat{B}$, or $\widehat{A} \Rightarrow B$ (resp., $A \Rightarrow \widetilde{B}$, $\widetilde{A} \Rightarrow \widetilde{B}$, or $\widetilde{A} \Rightarrow B$) [4]. Observe that the rule $A \Rightarrow B$ can be considered a specialized rule of the rule $\widehat{A \Rightarrow B}$.

It is known that a problem found in the data mining process is related to the huge quantity of patterns obtained, which complicates the user interpretation. To overcome this problem, many researchers use a variety of objective measures to evaluate the extracted knowledge to assist the user to understand and apply this knowledge ([16], [17], [18], [19], [20], [21], [22], [23], [24]). Thus, it is necessary to realize a study of the use of these measures in the knowledge evaluation considering the generalized rules context. Therefore, this paper realizes an analytical evaluation of some objective measures when applied in generalized association rules. It is important to do such evaluation since we can know beforehand the value of a measure in a generalized rule and if this value will be greater or less the value of the same measure in its specialized rules (see Section II). Thus, many relations were obtained among a generalized rule and its specialized rules in order to learn the behavior of some objective measures in the generalized rules context.

The paper is organized as follows. Section II presents an approach to obtain a generalized rules set in the post-processing step. This approach is presented once the analytical evaluation was realized considering that a generalized rule is obtained by grouping some specialized rules through taxonomies. Section III presents the analytical evaluation of the objective measures behavior when applied with generalized rules. Section IV presents a discussion of the obtained results. Finally, in Section V are the paper conclusions.

II. The Generalized Association Rule Post-Processing Approach (*GARPA*)

This paper considers the post-processing approach proposed by us in [15] to obtain a set of generalized association rules. Since there is an association rules set, obtained a priori with a traditional mining algorithm, the *GARPA* main idea, shown in Fig. 1, consists in generalizing this set based on a given domain taxonomy. The process obtains a generalized rules set composed by some rules that could not be generalized (for example, rule R40 shown in Fig. 1) and by some generalized rules obtained by grouping some rules (at least two rules) using the taxonomy set (for example, rule R35 shown in Fig. 1 – rule obtained by grouping the rules $milk_a \Rightarrow bread$ (R3), $milk_b \Rightarrow bread$ (R4) and $milk_c \Rightarrow bread$ (R7)). The generalization can be done in one side of the rule (antecedent (lhs: left hand side) or consequent (rhs: right hand side)) or in both sides ($lrhs$: left right hand side). Observe that a generalized item in a generalized rule is composed by the union of two or more specialized items contained in the taxonomy. Considering the notation previously presented, the input rules are here considered the specialized rules and the new rules, which contains a generalized item, the generalized rules.

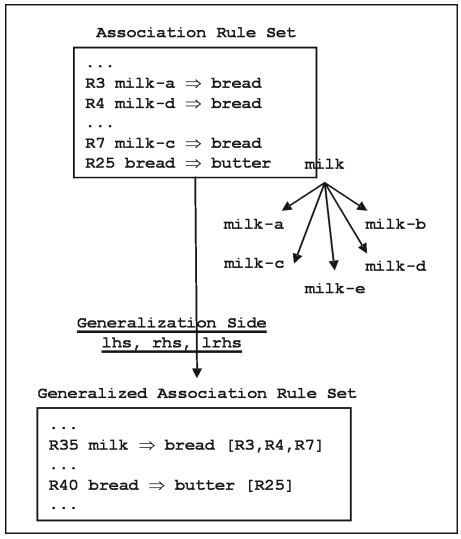

Fig. 1. The idea of *GARPA* approach.

III. The Objective Measures Analytical Evaluation

As mentioned before, the objective of this analytical evaluation is to study the objective measures behavior when applied in a generalized rule. To do such evaluation, for each of the measures considered in the study a relation was made among a generalized rule and its specialized rules, that is, the rules that were grouped by taxonomy to obtain the generalized rule (see Section II). As a generalized item can occur in both sides ($lrhs$) or in one side (lhs or rhs), the found relations were divided by side and are followed presented. The evaluation was made for some of the measures discussed in [21] and presented in Table I. A brief description of the measures meaning, including its references, can be found in [21]. It is important to mention that the measures shown in Table I were chosen due to its easy equation interpretation.

A. Measures Behavior in the LHS Generalization

In order to find the measures behavior in a generalized rule, considering a generalized item in the lhs side, the following assumption was done. Consider the rule $G \Rightarrow A$, obtained from the rules $B \Rightarrow A$ and $C \Rightarrow A$, where G is a generalized item formed by the grouping of the specialized items B and C. Thus, the event E_G is such that $E_G = E_B \cup E_C$. Based on this assumption, an analytical relation was found for each of the measures considered, which are followed presented.

TABLE I

OBJECTIVE MEASURES FOR AN ASSOCIATION PATTERN $A \Rightarrow B$ [21].

Measure	Definition	Symmetric
Added Value (AV)	$P(E_B\|E_A) - P(E_B)$	No
Certainty Factor (CF)	$\frac{P(E_B\|E_A)-P(E_B)}{1-P(E_B)}$	No
Confidence (Conf)	$P(E_B\|E_A)$	No
Conviction (Conv)	$\frac{P(E_A)P(\overline{E_B})}{P(E_A \cap \overline{E_B})}$	No
Laplace (L)	$\frac{\|D\|P(E_A \cap E_B)+1}{\|D\|P(E_A)+2}$	No
Interest (I)	$\frac{P(E_A \cap E_B)}{P(E_A)P(E_B)}$	Yes
Jaccard (ζ)	$\frac{P(E_A \cap E_B)}{P(E_A)+P(E_B)-P(E_A \cap E_B)}$	Yes
Piatetsky-Shapiro's (PS)	$P(E_A \cap E_B) - P(E_A)P(E_B)$	Yes
Support (Sup)	$P(E_A \cap E_B)$	Yes

1) Added Value: Applying the *Added Value* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$\begin{aligned}
AV_{G,A} &= P(E_A|E_G) - P(E_A) \\
&= P(E_A|E_B \cup E_C) - P(E_A) \\
&= \frac{P[E_A \cap (E_B \cup E_C)]}{P(E_B \cup E_C)} - P(E_A) \\
&= \frac{P[(E_A \cap E_B) \cup (E_A \cap E_C)]}{P(E_B \cup E_C)} - P(E_A) \\
&= \frac{P(E_A \cap E_B) + P(E_A \cap E_C) - P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)} - P(E_A) \\
&= [\frac{P(E_A \cap E_B)}{P(E_B \cup E_C)} - P(E_A)] + [\frac{P(E_A \cap E_C)}{P(E_B \cup E_C)} - P(E_A)] \\
&\quad - [\frac{P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)} - P(E_A)] \\
&\leqslant AV_{B,A} + AV_{C,A} - [\frac{P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)} - P(E_A)] \qquad (1)
\end{aligned}$$

2) Certainty Factor: Applying the *Certainty Factor* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$\begin{aligned}
CF_{G,A} &= \frac{P(E_A|E_G) - P(E_A)}{1 - P(E_A)} \\
&= \frac{P(E_A|E_B \cup E_C) - P(E_A)}{1 - P(E_A)} \\
&= \frac{\frac{P[E_A \cap (E_B \cup E_C)]}{P(E_B \cup E_C)} - P(E_A)}{1 - P(E_A)} \\
&= \frac{\frac{P(E_A \cap E_B) + P(E_A \cap E_C) - P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)} - P(E_A)}{1 - P(E_A)} \\
&= [\frac{\frac{P(E_A \cap E_B)}{P(E_B \cup E_C)} - P(E_A)}{1 - P(E_A)}] + [\frac{\frac{P(E_A \cap E_C)}{P(E_B \cup E_C)} - P(E_A)}{1 - P(E_A)}] \\
&\quad - [\frac{\frac{P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)} - P(E_A)}{1 - P(E_A)}] \\
&\leqslant CF_{B,A} + CF_{C,A} - [\frac{P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)P(\overline{E_A})} - \frac{P(E_A)}{P(\overline{E_A})}] \qquad (2)
\end{aligned}$$

3) Confidence: Applying the *Confidence* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$\begin{aligned}
Conf_{G,A} &= P(E_A|E_G) \\
&= P(E_A|E_B \cup E_C) \\
&= \frac{P[E_A \cap (E_B \cup E_C)]}{P(E_G)} \\
&= \frac{P(E_A \cap E_B)}{P(E_G)} + \frac{P(E_A \cap E_C)}{P(E_G)} - \frac{P(E_A \cap E_B \cap E_C)}{P(E_G)} \\
&= \frac{P(E_B)}{P(E_G)} \times Conf_{B,A} + \frac{P(E_C)}{P(E_G)} \times Conf_{C,A} \\
&\quad - \frac{P(E_B \cap E_C)}{P(E_G)} \times Conf_{B \cap C,A} \\
&\leqslant Conf_{B,A} + Conf_{C,A} - \frac{P(E_B \cap E_C)}{P(E_B \cup E_C)} \times Conf_{B \cap C,A} \qquad (3)
\end{aligned}$$

4) Conviction: Applying the *Conviction* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$\begin{aligned}
Conv_{G,A} &= \frac{P(E_G)P(\overline{E_A})}{P(E_G \cap \overline{E_A})} \\
&= \frac{P(E_B \cup E_C)P(\overline{E_A})}{P(E_G \cap \overline{E_A})} \\
&= \frac{P(E_B)P(\overline{E_A})}{P(E_G \cap \overline{E_A})} + \frac{P(E_C)P(\overline{E_A})}{P(E_G \cap \overline{E_A})} - \frac{P(E_B \cap E_C)P(\overline{E_A})}{P(E_G \cap \overline{E_A})} \\
&= \frac{P(E_B \cap \overline{E_A})}{P(E_G \cap \overline{E_A})} \times Conv_{B,A} + \frac{P(E_C \cap \overline{E_A})}{P(E_G \cap \overline{E_A})} \times Conv_{C,A} \\
&\quad - \frac{P(E_B \cap E_C \cap \overline{E_A})}{P(E_G \cap \overline{E_A})} \times Conv_{B \cap C,A} \\
&\leqslant Conv_{B,A} + Conv_{C,A} \\
&\quad - \frac{P(E_B \cap E_C \cap \overline{E_A})}{P[(E_B \cup E_C) \cap \overline{E_A}]} \times Conv_{B \cap C,A} \qquad (4)
\end{aligned}$$

5) Laplace: Applying the *Laplace* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$
\begin{aligned}
L_{G,A} &= \frac{|D|P(E_G \cap E_A)+1}{|D|P(E_G)+2} \\
&= \frac{|D|P[(E_B \cup E_C) \cap E_A]+1}{|D|P(E_G)+2} \\
&= \frac{|D|P(E_B \cap E_A)+1}{|D|P(E_G)+2} + \frac{|D|P(E_C \cap E_A)+1}{|D|P(E_G)+2} \\
&\quad - \frac{|D|P(E_B \cap E_C \cap E_A)+1}{|D|P(E_G)+2} \\
&= \frac{|D|P(E_B)+2}{|D|P(E_G)+2} \times L_{B,A} + \frac{|D|P(E_C)+2}{|D|P(E_G)+2} \times L_{C,A} \\
&\quad - \frac{|D|P(E_B \cap E_C)+2}{|D|P(E_G)+2} \times L_{B \cap C,A} \\
&\leqslant L_{B,A} + L_{C,A} - \frac{|D|P(E_B \cap E_C)+2}{|D|P(E_B \cup E_C)+2} \times L_{B \cap C,A}
\end{aligned} \tag{5}
$$

6) Interest: Applying the *Interest* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$
\begin{aligned}
I_{G,A} &= \frac{P(E_G \cap E_A)}{P(E_G)P(E_A)} \\
&= \frac{P[(E_B \cup E_C) \cap E_A]}{P(E_G)P(E_A)} \\
&= \frac{P(E_B \cap E_A)}{P(E_G)P(E_A)} + \frac{P(E_C \cap E_A)}{P(E_G)P(E_A)} - \frac{P(E_B \cap E_C \cap E_A)}{P(E_G)P(E_A)} \\
&= \frac{P(E_B)P(E_A)}{P(E_G)P(E_A)} \times I_{B,A} + \frac{P(E_C)P(E_A)}{P(E_G)P(E_A)} \times I_{C,A} \\
&\quad - \frac{P(E_B \cap E_C)P(E_A)}{P(E_G)P(E_A)} \times I_{B \cap C,A} \\
&\leqslant I_{B,A} + I_{C,A} - \frac{P(E_B \cap E_C)}{P(E_B \cup E_C)} \times I_{B \cap C,A}
\end{aligned} \tag{6}
$$

7) Jaccard: Applying the *Jaccard* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$
\begin{aligned}
\zeta_{G,A} &= \frac{P(E_G \cap E_A)}{P(E_G)+P(E_A)-P(E_G \cap E_A)} \\
&= \frac{P(E_G \cap E_A)}{P(E_G \cup E_A)} \\
&= \frac{P[(E_B \cup E_C) \cap E_A]}{P(E_G \cup E_A)} \\
&= \frac{P(E_B \cap E_A)}{P(E_G \cup E_A)} + \frac{P(E_C \cap E_A)}{P(E_G \cup E_A)} - \frac{P(E_B \cap E_C \cap E_A)}{P(E_G \cup E_A)} \\
&= \frac{P(E_B \cup E_A)}{P(E_G \cup E_A)} \times \zeta_{B,A} + \frac{P(E_C \cup E_A)}{P(E_G \cup E_A)} \times \zeta_{C,A} \\
&\quad - \frac{P[(E_B \cap E_C) \cup E_A]}{P(E_G \cup E_A)} \times \zeta_{B \cap C,A} \\
&\leqslant \zeta_{B,A} + \zeta_{C,A} - \frac{P[(E_B \cap E_C) \cup E_A]}{P(E_B \cup E_C \cup E_A)} \times \zeta_{B \cap C,A}
\end{aligned} \tag{7}
$$

8) Piatetsky-Shapiro's: Applying the *Piatetsky-Shapiro's* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$
\begin{aligned}
PS_{G,A} &= P(E_G \cap E_A) - P(E_G)P(E_A) \\
&= P[(E_B \cup E_C) \cap E_A] - P(E_B \cup E_C)P(E_A) \\
&= P[(E_B \cap E_A) \cup (E_C \cap E_A)] \\
&\quad - [P(E_B)+P(E_C)-P(E_B \cap E_C)]P(E_A) \\
&= [P(E_B \cap E_A)+P(E_C \cap E_A)-P(E_B \cap E_C \cap E_A)] \\
&\quad - [P(E_B)P(E_A)+P(E_C)P(E_A)-P(E_B \cap E_C)P(E_A)] \\
&= P(E_B \cap E_A) - P(E_B)P(E_A) + P(E_C \cap E_A) - P(E_C)P(E_A) \\
&\quad - P(E_B \cap E_C \cap E_A) + P(E_B \cap E_C)P(E_A) \\
&= PS_{B,A} + PS_{C,A} - PS_{B \cap C,A}
\end{aligned} \tag{8}
$$

9) Support: Applying the *Support* measure to the rule $G \Rightarrow A$, the following relation is obtained.

$$
\begin{aligned}
Sup_{G,A} &= P(E_G \cap E_A) \\
&= P[(E_B \cup E_C) \cap E_A] \\
&= P[(E_B \cap E_A) \cup (E_C \cap E_A)] \\
&= P(E_B \cap E_A) + P(E_C \cap E_A) - P(E_B \cap E_C \cap E_A) \\
&= Sup_{B,A} + Sup_{C,A} - Sup_{B \cap C,A} \\
&\quad Since \\
&\qquad Sup_{B,A} \geqslant Sup_{B \cap C,A}; \\
&\qquad Sup_{C,A} \geqslant Sup_{B \cap C,A} \\
&\qquad \Rightarrow Sup_{G,A} \geqslant \max\{Sup_{B,A}; Sup_{C,A}\}
\end{aligned} \tag{9}
$$

B. Measures Behavior in the RHS Generalization

In order to find the measures behavior in a generalized rule, considering a generalized item in the rhs side, the following assumption was done. Consider the rule $A \Rightarrow G$, obtained from the rules $A \Rightarrow B$ and $A \Rightarrow C$, where G is a generalized item formed by the grouping of the specialized items B and C. Thus, the event E_G is such that $E_G = E_B \cup E_C$. Based on this assumption, an analytical relation was found for each of the measures considered, which are followed presented.

1) Added Value: Applying the *Added Value* measure to the rule $A \Rightarrow G$, the following relation is obtained.

$$
\begin{aligned}
AV_{A,G} &= P(E_G|E_A) - P(E_G) \\
&= P(E_B \cup E_C|E_A) - P(E_B \cup E_C) \\
&= [P(E_B|E_A)+P(E_C|E_A)-P(E_B \cap E_C|E_A)] \\
&\quad - [P(E_B)+P(E_C)-P(E_B \cap E_C)] \\
&= [P(E_B|E_A)-P(E_B)]+[P(E_C|E_A)-P(E_C)] \\
&\quad - [P(E_B \cap E_C|E_A)-P(E_B \cap E_C)] \\
&= AV_{A,B} + AV_{A,C} - AV_{A,B \cap C}
\end{aligned} \tag{10}
$$

2) Certainty Factor: Applying the *Certainty Factor* measure to the rule $A \Rightarrow G$, the following relation is obtained.

$$
\begin{aligned}
CF_{A,G} &= \frac{P(E_G|E_A)-P(E_G)}{1-P(E_G)} \\
&= \frac{P(E_B \cup E_C|E_A)-P(E_B \cup E_C)}{1-P(E_G)} \\
&= \frac{P(E_B|E_A)}{P(\overline{E_G})} + \frac{P(E_C|E_A)}{P(\overline{E_G})} - \frac{P(E_B \cap E_C|E_A)}{P(\overline{E_G})} \\
&\quad - \frac{P(E_B)}{P(\overline{E_G})} - \frac{P(E_C)}{P(\overline{E_G})} + \frac{P(E_B \cap E_C)}{P(\overline{E_G})} \\
&= \frac{P(\overline{E_B})}{P(\overline{E_G})} \times CF_{A,B} + \frac{P(\overline{E_C})}{P(\overline{E_G})} \times CF_{A,C} \\
&\quad - \frac{P(\overline{E_B} \cup \overline{E_C})}{P(\overline{E_G})} \times CF_{A,B \cap C} \\
&\geqslant CF_{A,B} + CF_{A,C} - \frac{P(\overline{E_B} \cup \overline{E_C})}{P(\overline{E_B} \cap \overline{E_C})} \times CF_{A,B \cap C}
\end{aligned} \tag{11}
$$

3) Confidence: Applying the *Confidence* measure to the rule $A \Rightarrow G$, the following relation is obtained.

$$
\begin{aligned}
Conf_{A,G} &= P(E_G|E_A) \\
&= P(E_B \cup E_C|E_A) \\
&= P(E_B|E_A) + P(E_C|E_A) - P(E_B \cap E_C|E_A) \\
&= Conf_{A,B} + Conf_{A,C} - Conf_{A,B\cap C} \\
Since& \\
&Conf_{A,B} \geqslant Conf_{A,B\cap C}; \\
&Conf_{A,C} \geqslant Conf_{A,B\cap C} \\
&\Rightarrow Conf_{A,G} \geqslant \max\{Conf_{A,B}; Conf_{A,C}\}
\end{aligned} \tag{12}
$$

4) Conviction: Applying the *Conviction* measure to the rule $A \Rightarrow G$, the following relation is obtained.

$$
\begin{aligned}
Conv_{A,G} &= \frac{P(E_A)P(\overline{E_G})}{P(E_A \cap \overline{E_G})} \\
&= \frac{P(E_A)P(\overline{E_B \cup E_C})}{P[E_A \cap (\overline{E_B \cup E_C})]} \\
&= \frac{P(E_A)P(\overline{E_B} \cap \overline{E_C})}{P(E_A \cap \overline{E_B} \cap \overline{E_C})} \\
Since& \\
&\frac{P(E_A)P(\overline{E_B} \cap \overline{E_C})}{P(E_A \cap \overline{E_B} \cap \overline{E_C})} \geqslant \frac{P(E_A)P(\overline{E_B} \cap \overline{E_C})}{P(E_A \cap \overline{E_B})}; \\
&\frac{P(E_A)P(\overline{E_B} \cap \overline{E_C})}{P(E_A \cap \overline{E_B} \cap \overline{E_C})} \geqslant \frac{P(E_A)P(\overline{E_B} \cap \overline{E_C})}{P(E_A \cap \overline{E_C})} \\
&\Rightarrow Conv_{A,G} \geqslant \frac{P(\overline{E_B} \cap \overline{E_C})}{P(\overline{E_B})} \times Conv_{A,B}; \\
&Conv_{A,G} \geqslant \frac{P(\overline{E_B} \cap \overline{E_C})}{P(\overline{E_C})} \times Conv_{A,C}
\end{aligned} \tag{13}
$$

5) Laplace: Applying the *Laplace* measure to the rule $A \Rightarrow G$, the following relation is obtained.

$$
\begin{aligned}
L_{A,G} &= \frac{|D|P(E_A \cap E_G) + 1}{|D|P(E_A) + 2} \\
&= \frac{|D|P[E_A \cap (E_B \cup E_C)] + 1}{|D|P(E_A) + 2} \\
&= \frac{|D|P[(E_A \cap E_B) \cup (E_A \cap E_C)] + 1}{|D|P(E_A) + 2} \\
&= \frac{|D|[P(E_A \cap E_B) + P(E_A \cap E_C) - P(E_A \cap E_B \cap E_C)] + 1}{|D|P(E_A) + 2} \\
&= \frac{|D|P(E_A \cap E_B) + 1}{|D|P(E_A) + 2} + \frac{|D|P(E_A \cap E_C) + 1}{|D|P(E_A) + 2} \\
&\quad - \frac{|D|P(E_A \cap E_B \cap E_C) + 1}{|D|P(E_A) + 2} \\
&= L_{A,B} + L_{A,C} - L_{A,B\cap C} \\
Since& \\
&L_{A,B} \geqslant L_{A,B\cap C}; \\
&L_{A,C} \geqslant L_{A,B\cap C} \\
&\Rightarrow L_{A,G} \geqslant \max\{L_{A,B}; L_{A,C}\}
\end{aligned} \tag{14}
$$

6) Interest: Since the *Interest* measure is symmetric, only the obtained relation to the rule $A \Rightarrow G$ is shown.

$$
I_{A,G} \leqslant I_{A,B} + I_{A,C} - \frac{P(E_B \cap E_C)}{P(E_B \cup E_C)} \times I_{A,B\cap C} \tag{15}
$$

7) Jaccard: Since the *Jaccard* measure is symmetric, only the obtained relation to the rule $A \Rightarrow G$ is shown.

$$
\zeta_{A,G} \leqslant \zeta_{A,B} + \zeta_{A,C} - \frac{P[E_A \cup (E_B \cap E_C)]}{P(E_A \cup E_B \cup E_C)} \times \zeta_{A,B\cap C} \tag{16}
$$

8) Piatetsky-Shapiro's: Since the *Piatetsky-Shapiro's* measure is symmetric, only the obtained relation to the rule $A \Rightarrow G$ is shown.

$$
PS_{A,G} = PS_{A,B} + PS_{A,C} - PS_{A,B\cap C} \tag{17}
$$

9) Support: Since the *Support* measure is symmetric, only the obtained relation to the rule $A \Rightarrow G$ is shown.

$$
Sup_{A,G} \geqslant \max\{Sup_{A,B}; Sup_{A,C}\} \tag{18}
$$

C. Measures Behavior in the $LRHS$ Generalization

In order to find the measures behavior in a generalized rule, considering generalized items in both sides, the following assumption was done. Consider the rule $G_1 \Rightarrow G_2$, obtained from the rules $A \Rightarrow C$ and $B \Rightarrow D$, where G_1 is a generalized item formed by the grouping of the specialized items A and B and G_2 a generalized item formed by the grouping of the specialized items C and D. Thus, the event E_{G_1} is such that $E_{G_1} = E_A \cup E_B$ and E_{G_2} is such that $E_{G_2} = E_C \cup E_D$. Based on this assumption, an analytical relation was found for each of the measures considered, which are followed presented.

1) Added Value: In order to obtain the $G_1 \Rightarrow G_2$ relation to the *Added Value* measure, we apply (1) in the current rule.

$$
\begin{aligned}
AV_{G_1,G_2} \leqslant\ & AV_{A,C\cup D} + AV_{B,C\cup D} \\
& - [\frac{P[(E_C \cup E_D) \cap E_A \cap E_B]}{P(E_A \cup E_B)} - P(E_C \cup E_D)]
\end{aligned} \tag{19}
$$

Now, applying (10) in (19) we have:

$$
\begin{aligned}
AV_{G_1,G_2} \leqslant\ & AV_{A,C} + AV_{A,D} - AV_{A,C\cap D} \\
& + AV_{B,C} + AV_{B,D} - AV_{B,C\cap D} \\
& - [\frac{P[(E_C \cup E_D) \cap E_A \cap E_B]}{P(E_A \cup E_B)} - P(E_C \cup E_D)]
\end{aligned} \tag{20}
$$

Since the same idea is used to obtain all the others measures relations in the $lrhs$ generalization, the mathematics deductions are not shown.

IV. Discussion of the Evaluation Results

Tables II, III, and IV show the objective measures behaviors through the relations found for each of the measures in each side. These relations represent the existing connection among the value of a measure in one generalized rule and the values of the same measure in its specialized rules.

In order to demonstrate how to interpret the found relations, consider the *Support* measure. One of the relations found to this measure is $Sup_{G,A} \geqslant \max\{Sup_{B,A}; Sup_{C,A}\}$,

TABLE II
OBJECTIVE MEASURES RELATIONS IN THE LHS GENERALIZATION.

Measure	$G \Rightarrow A$ ($G = B \cup C$)
Added Value (AV)	$AV_{G,A} \leqslant AV_{B,A} + AV_{C,A} - [\frac{P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)} - P(E_A)]$
Certainty Factor (CF)	$CF_{G,A} \leqslant CF_{B,A} + CF_{C,A} - [\frac{P(E_A \cap E_B \cap E_C)}{P(E_B \cup E_C)P(\overline{E_A})} - \frac{P(E_A)}{P(\overline{E_A})}]$
Confidence (Conf)	$Conf_{G,A} \leqslant Conf_{B,A} + Conf_{C,A} - \frac{P(E_B \cap E_C)}{P(E_B \cup E_C)} \times Conf_{B \cap C,A}$
Conviction (Conv)	$Conv_{G,A} \leqslant Conv_{B,A} + Conv_{C,A} - \frac{P(E_B \cap E_C \cap \overline{E_A})}{P[(E_B \cup E_C) \cap \overline{E_A}]} \times Conv_{B \cap C,A}$
Laplace (L)	$L_{G,A} \leqslant L_{B,A} + L_{C,A} - \frac{\lvert D \rvert P(E_B \cap E_C)+2}{\lvert D \rvert P(E_B \cup E_C)+2} \times L_{B \cap C,A}$
Interest (I)	$I_{G,A} \leqslant I_{B,A} + I_{C,A} - \frac{P(E_B \cap E_C)}{P(E_B \cup E_C)} \times I_{B \cap C,A}$
Jaccard (ζ)	$\zeta_{G,A} \leqslant \zeta_{B,A} + \zeta_{C,A} - \frac{P[(E_B \cap E_C) \cup E_A]}{P(E_B \cup E_C \cup E_A)} \times \zeta_{B \cap C,A}$
Piatetsky-Shapiro's (PS)	$PS_{G,A} = PS_{B,A} + PS_{C,A} - PS_{B \cap C,A}$
Support (Sup)	$Sup_{G,A} \geqslant \max\{Sup_{B,A}; Sup_{C,A}\}$

TABLE III
OBJECTIVE MEASURES RELATIONS IN THE RHS GENERALIZATION.

Measure	$A \Rightarrow G$ ($G = B \cup C$)
Added Value (AV)	$AV_{A,G} = AV_{A,B} + AV_{A,C} - AV_{A,B \cap C}$
Certainty Factor (CF)	$CF_{A,G} \geqslant CF_{A,B} + CF_{A,C} - \frac{P(\overline{E_B} \cup \overline{E_C})}{P(\overline{E_B} \cap \overline{E_C})} \times CF_{A,B \cap C}$
Confidence (Conf)	$Conf_{A,G} \geqslant \max\{Conf_{A,B}; Conf_{A,C}\}$
Conviction (Conv)	$Conv_{A,G} \geqslant \frac{P(\overline{E_B} \cap \overline{E_C})}{P(\overline{E_B})} \times Conv_{A,B}$; $\frac{P(\overline{E_B} \cap \overline{E_C})}{P(\overline{E_C})} \times Conv_{A,C}$
Laplace (L)	$L_{A,G} \geqslant \max\{L_{A,B}; L_{A,C}\}$
Interest (I)	$I_{A,G} \leqslant I_{A,B} + I_{A,C} - \frac{P(E_B \cap E_C)}{P(E_B \cup E_C)} \times I_{A,B \cap C}$
Jaccard (ζ)	$\zeta_{A,G} \leqslant \zeta_{A,B} + \zeta_{A,C} - \frac{P[E_A \cup (E_B \cap E_C)]}{P(E_A \cup E_B \cup E_C)} \times \zeta_{A,B \cap C}$
Piatetsky-Shapiro's (PS)	$PS_{A,G} = PS_{A,B} + PS_{A,C} - PS_{A,B \cap C}$
Support (Sup)	$Sup_{A,G} \geqslant \max\{Sup_{A,B}; Sup_{A,C}\}$

which means that the support value of a generalized rule ($Sup_{G,A}$), containing a generalized item in the lhs side, is greater or equal to the maximum *Sup* value of its specialized rules ($Sup_{B,A}$ and $Sup_{C,A}$). Considering another measure, *Piatetsky-Shapiro's*, for example, we can find the following relation: $PS_{A,G} = PS_{A,B} + PS_{A,C} - PS_{A,B \cap C}$, which means that the *PS* value of a generalized rule ($PS_{A,G}$), containing a generalized item in the rhs side, is equal to the sum of the *PS* values of its specialized rules ($PS_{A,B}$ and $PS_{A,C}$) minus the *PS* value considering the intersection of the specialized items used to form the generalized item ($PS_{A,B \cap C}$). The interpretation of the rest of the measures follows the same idea. In relation to the $lrhs$, observe in Table IV that we did not arrive at a conclusion for some measures (*CF*, *Conf*, *Conv*, and *L*), since the behavior in all of these measures in the lhs presented a $\leqslant$ inequality relation and in the rhs a $\geqslant$ inequality relation.

From Tables II, III and IV it can be observed that some measures are written by an inequality to express the value of a generalized rule. This inequality represents the measure behavior in relation to the specialized items used to form the generalized item. So, depending on the inequality direction it can be concluded that the behavior of a measure in a generalized rule will be equal or better in relation to the behavior of its specialized items ($\geqslant$ inequality) or that the behavior of a measure in a generalized rule will be equal or worse in relation to the behavior of its specialized items ($\leqslant$ inequality). So, it is suitable to use a measure that presents

TABLE IV
OBJECTIVE MEASURES RELATIONS IN THE $LRHS$ GENERALIZATION.

Measure	$G_1 \Rightarrow G_2$ ($G_1 = A \cup B$; $G_2 = C \cup D$)
Added Value (AV)	$AV_{G_1,G_2} \leqslant AV_{A,C} + AV_{A,D} - AV_{A,C\cap D} + AV_{B,C} + AV_{B,D} - AV_{B,C\cap D}$ $-[\frac{P[(E_C\cup E_D)\cap E_A\cap E_B]}{P(E_A\cup E_B)} - P(E_C \cup E_D)]$
Certainty Factor (CF)	No Conclusion
Confidence (Conf)	No Conclusion
Conviction (Conv)	No Conclusion
Laplace (L)	No Conclusion
Interest (I)	$I_{G_1,G_2} \leqslant I_{A,C} + I_{A,D} - \frac{P(E_C\cap E_D)}{P(E_C\cup E_D)} \times I_{A,C\cap D} + I_{B,C} + I_{B,D} - \frac{P(E_C\cap E_D)}{P(E_C\cup E_D)} \times I_{B,C\cap D}$ $-\frac{P(E_A\cap E_B)}{P(E_A\cup E_B)}[I_{A\cap B,C} + I_{A\cap B,D} - \frac{P(E_C\cap E_D)}{P(E_C\cup E_D)} \times I_{A\cap B,C\cap D}]$
Jaccard (ζ)	$\zeta_{G1,G_2} \leqslant \zeta_{A,C} + \zeta_{A,D} - \frac{P[E_A\cup(E_C\cap E_D)]}{P(E_A\cup E_C\cup E_D)} \times \zeta_{A,C\cap D} + \zeta_{B,C} + \zeta_{B,D} - \frac{P[E_B\cup(E_C\cap E_D)]}{P(E_B\cup E_C\cup E_D)} \times \zeta_{B,C\cap D}$ $-\frac{P[(E_A\cap E_B)\cup E_C\cup E_D]}{P(E_A\cup E_B\cup E_C\cup E_D)}[\zeta_{A\cap B,C} + \zeta_{A\cap B,D} - \frac{P[(E_A\cap E_B)\cup(E_C\cap E_D)]}{P[(E_A\cap E_B)\cup E_C\cup E_D]} \times \zeta_{A\cap B,C\cap D}]$
Piatetsky-Shapiro's (PS)	$PS_{G_1,G_2} = PS_{A,C} + PS_{A,D} - PS_{A,C\cap D} + PS_{B,C} + PS_{B,D} - PS_{B,C\cap D}$ $-PS_{A\cap B,C} - PS_{A\cap B,D} + PS_{A\cap B,C\cap D}$
Support (Sup)	$Sup_{G_1,G2} \geqslant \max\{Sup_{A,C}; Sup_{A,D}; Sup_{B,C}; Sup_{B,D}\}$

an improvement in its behavior compared to its specialized items in order to select the most interesting rules to the user. Therefore, note that are some measures that are better to be used in the selection of rules that contain a generalized item in one specific side, as *CF*, *Conf*, *Conv* and *L*, that presents a better behavior when applied in the rhs side.

It is also interesting to note that analyzing Table I the following relations can be obtained: $CF = \frac{AV}{1-P(E_B)}$, $Conf = AV + P(E_B)$, $I = \frac{AV}{P(E_B)} + 1$ and $Conv = \frac{P(E_B)}{1-Conf}$. This means that when we are analyzing a generalized rule that contains a generalized item in the lhs, the *CF*, *Conf*, I and *Conv* measures present the same behavior, since they are expressed in terms of *AV*.

Finally, it can be noted that the *Support* and the *Confidence* measures presented an expected behavior (Section I). On despite of this, an analysis of the other measures behavior, when applied to generalized rules, has not been found, which represents a relevant contribution. So, the main advantages of this analytical evaluation are:

[a] to know beforehand the behavior of an objective measure when applied in a generalized rule;
[b] to compute more easily the objective measures values for a generalized rule since we can use the found relations to obtain these values;
[c] to provide an overview of the objective measures behavior to the data mining community when applied in a generalized rule.

It is important to bounce that this analytical evaluation is valid, independent of the data mining step where the generalized rules are obtained, to any generalized rules that contains a generalized item that can be viewed as the union of two or more specialized items contained in a given taxonomy.

V. CONCLUSION

Generalized association rules are rules that contain some background knowledge giving a more general view of the domain. This knowledge is commonly codified by a taxonomy set over the data set items. Many researches use taxonomies in different data mining steps to obtain generalized rules. However, a problem identified with the techniques used to obtain these types of patterns is the amount of rules obtained, since the objective of the data mining process is to obtain a useful and interesting knowledge to support the user's decisions. To help the users to the select these pieces of knowledge there are many objective measures. These measures have been studied in many types of rules. However, an analysis of the behavior of these measures, when applied to evaluate generalized association rules, has not been found. In this context, this paper presents an analytical evaluation over the behavior of some objective measures to verify the existing relation among an objective measure value in a generalized rule and an objective measure value in its specialized rules.

Tables II, III, and IV present all the valid relations found for each of the measures analyzed considering a generalized item in both of the sides of a rule and in a specific side of a rule. It is important to highlight that these relations represent the behavior of each of these measures in the context of a generalized rule, which is a meaningful contribution of this paper. As a continuation of this work, other objective measures not considered in the analytical evaluation will be studied with the purpose of understanding its behavior in the generalized association rules context.

It is important to mention that in [25] an experimental evaluation of the presented measures was done and a grouping measure was generated according to the generalization side. These measure groups are useful to help the specialists to choose an appropriate measure to evaluate their generalized

rules. So, this current work is, in fact, a complementary study of [25]. As a future research of these works, an analysis with an expert domain will be done to verify if the measure groups are, in fact, able to select the most interesting generalized rules depending on the generalization side.

ACKNOWLEDGMENT

We wish to thank the Instituto Fábrica do Milênio (IFM) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the financial support.

REFERENCES

[1] R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994, 487–499.

[2] J. Hipp and U. Güntzer and G. Nakhaeizadeh, Data Mining of Association Rules and the Process of Knowledge Discovery in Databases, *Lecture Notes in Artificial Intelligence – Advances in Data Mining*, 2394, 2002, 15–36.

[3] R. Srikant and R. Agrawal, Mining Generalized Association Rules, *Proceedings of the 21st International Conference on Very Large Data Bases*, 1995, 407–419.

[4] J.-M. Adamo, *Data Mining for Association Rules and Sequential Patterns* (Springer-Verlag, 2001).

[5] R. Srikant and R. Agrawal, Mining Generalized Association Rules, *Future Generation Computer Systems*, 13(2/3), 1997, 161–180.

[6] J. Hipp and A. Myka and R. Wirth and U. Güntzer, A New Algorithm for Faster Mining of Generalized Association Rules, *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery*, 1998, 74–82.

[7] I. Weber, On pruning Strategies for Discovery of Generalized and Quantitative Association Rules, *Proceedings of the Knowledge Discovery and Data Mining Workshop*, 1998, 8 pp.

[8] J. Baixeries and G. Casas and J. L. Balcázar, *Frequent Sets, Sequences, and Taxonomies: New, Efficient Algorithmic Proposals* (Departament de LSI – Universitat Politècnica de Catalunya, LSI-00-78-R, 2000).

[9] S.-J. Yen and A. L. P. Chen, A Graph-Based Approach for Discovering Various Types of Association Rules, *IEEE Transactions on Knowledge and Data Engineering*, 13(5), 2001, 839–845.

[10] K. Sriphaew and T. Theeramunkong, Fast Algorithms for Mining Generalized Frequent Patterns of Generalized Association Rules, *IEICE Transactions on Information and Systems*, 87(3), 2004, 761–770.

[11] J. Han and Y. Fu, Mining Multiple-level Association Rules in Large Databases, *IEEE Transactions on Knowledge and Data Engineering*, 11(5), 1999, 798–805.

[12] F. Chung and C. Lui, A Post-analysis Framework for Mining Generalized Association Rules with Multiple Minimum Supports, *Proceedings of the Post-processing in Machine Learning and Data Mining: Interpretation, Visualization, Integration, and Related Topics (Workshop within KDD'2000)*, 2000.

[13] Y.-F. Huang and C.-M. Wu, Mining Generalized Association Rules Using Pruning Techniques, *Proceedings of the 2002 IEEE International Conference on Data Mining*, 2002, 227–234.

[14] M. A. Domingues and S. O. Rezende, Using Taxonomies to Facilitate the Analysis of the Association Rules, *Proceedings of ECML/PKDD'05 – The 2nd International Workshop on Knowledge Discovery and Ontologies*, 2005, 59–66.

[15] V. O. Carvalho and S. O. Rezende and M. de Castro, Regras de Associação Generalizadas: Obtenção e Avaliação, *Proceedings of the II Workshop em Algoritmos e Aplicações de Mineração de Dados*, 2006, 81–88.

[16] B. Liu and W. Hsu and S. Chen and Y. Ma, Analyzing the Subjective Interestingness of Association Rules, *Intelligent Systems and Their Applications*, 15(5), 2000, 47–55.

[17] R. J. Hilderman and H. J. Hamilton, Evaluation of Interestingness Measures for Ranking Discovered Knowledge, *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2001, 247–259.

[18] E. R. Omiecinski, Alternative Interest Measures for Mining Associations in Databases, *IEEE Transactions on Knowledge and Data Engineering*, 15(1), 2003, 57–69.

[19] E. A. Melanda and S. O. Rezende, Combining Quality Measures to Identify Interesting Association Rules, *Proceedings of the IX Ibero-American Conference on Artificial Intelligence*, 3315, 2004, 441–453.

[20] L. M. Sheikh and B. Tanveer and M. A. Hamdani, Interesting Measures for Mining Association Rules, *Proceedings of INMIC 2004 – 8th International*, 2004, 641–644.

[21] P.-N. Tan and V. Kumar and J. Srivastava, Selecting the Right Objective Measure for Association Analysis, *Information Systems*, 29(4), 2004, 293–313.

[22] R. Natarajan and B. Shekar, A Relatedness-Based Data-Driven Approach to Determination of Interestingness of Association Rules, *Proceedings of the 2005 ACM symposium on Applied computing*, 2005, 551–552.

[23] D. R. Carvalho and A. A. Freitas and N. Ebecken, Evaluating the Correlation between Objective Rule Interestingness Measures and Real Human Interest, *Proceedings of the Knowledge Discovery in Databases – PKDD-2005*, 2005, 453–461.

[24] R. Tamir and Y. Singer, On a Confidence Gain Measure for Association Rule Discovery and Scoring, *The VLDB Journal The International Journal on Very Large Data Bases*, 15(1), 2006, 40–52.

[25] V. O. Carvalho and S. O. Rezende and M. de Castro, Evaluating Generalized Association Rules through Objective Measures, *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications – AIA 2007*, 2007, in press.