Metrics to Support the Evaluation of Association Rule Clustering

Veronica Oliveira de Carvalho[1], Fabiano Fernandes dos Santos[2], and Solange Oliveira Rezende[2]

[1] Instituto de Geociências e Ciências Exatas,
UNESP - Univ Estadual Paulista, Rio Claro, Brazil
veronica@rc.unesp.br
[2] Instituto de Ciências Matemáticas e de Computação,
USP - Universidade de São Paulo, São Carlos, Brazil
{fabianof,solange}@icmc.usp.br

Abstract. Many topics related to association mining have received attention in the research community, especially the ones focused on the discovery of interesting knowledge. A promising approach, related to this topic, is the application of clustering in the pre-processing step to aid the user to find the relevant associative patterns of the domain. In this paper, we propose nine metrics to support the evaluation of this kind of approach. The metrics are important since they provide criteria to: (a) analyze the methodologies, (b) identify their positive and negative aspects, (c) carry out comparisons among them and, therefore, (d) help the users to select the most suitable solution for their problems. Some experiments were done in order to present how the metrics can be used and their usefulness.

Keywords: Association Rules, Clustering, Pre-processing.

1 Introduction

In the last years, researches have adopted some strategies to aid the user to find the relevant associative patterns of the domain. One of these strategies is to pre-process the data before obtaining the rules. For that, many approaches have been proposed, being clustering a promising one. In this case, the data are initially grouped into n groups. Association rules are extracted within each group and, in the end, n groups of rules are obtained. All these rules compose the rule set. According to [1], each group expresses its own associations without the interference of the other groups that contain different association patterns. The aim is to obtain potentially interesting rules that would not be extracted from unpartitioned data sets. The user must set the minimum support to a low value to discover these same patterns from unpartitioned data sets, causing a rapidly increase in the number of rules.

Distinct methodologies have been proposed to enable the described process. Each methodology uses a different combination of similarity measures with clustering algorithms to obtain the groups of rules. However, little has been done to

L. Bellatreche and M.K. Mohania (Eds.): DaWaK 2013, LNCS 8057, pp. 248–259, 2013.

analyze the performance of the methodologies or even to compare the results. So, there are some issues that have to be investigated:

Issue 1. Is there overlap between a rule set obtained through partitioned data, i.e., extracted from clustered data, in relation to a rule set obtained through unpartitioned data, i.e., extracted from traditional process? A ***r***ule ***s***et obtained through a ***p***artitioned data is named here as RsP and a ***r***ule ***s***et obtained through a ***t***raditional process is named here as RsT.

Issue 2. Is there overlap between the rules in RsT and RsP regarding the interesting knowledge? In other words, has RsP, in fact, more interesting patterns than RsT?

Issue 3. What is the process behavior regarding the number of rules that are obtained in RsP?

Based on the exposed arguments and on the three presented issues, nine metrics are proposed in this paper to support the evaluation of the methodologies that use clustering in the pre-processing step. Thereby, this paper will contribute with future researches since the metrics will provide criteria to: (a) analyze the methodologies, (b) identify their positive and negative aspects, (c) carry out comparisons among them and, therefore, (d) help the users to select the most suitable solution for their problems. It is important to say that the aim here is not to discover interesting rules, but to provide a standardized assessment procedure to support the evaluation of the methodologies that use clustering in the pre-processing step in order to discover the interesting rules.

This paper is organized as follows. Section 2 presents the proposed metrics. Section 3 describes some experiments that were carried out to show how the metrics can be used. Section 4 discusses the results obtained in the experiments. Section 5 surveys the related researches. Finally, conclusion is given in Section 6.

2 Proposed Evaluation Metrics

Nine metrics are proposed to support the evaluation of the methodologies that use clustering in the pre-processing step, as the methodologies described in Section 5. Each metric is related to an issue mentioned in Section 1. For each issue there are one or more metrics. To propose the metrics, we assume that RsP is better than RsT when it generates new knowledge in a few groups.

All metrics, with exception to M_{NR-RsP}, range from 0 to 1. Since RsP contains all the rules extracted within each group, repeated rules may exist in the set. In RsT the same doesn't occur since the rules are unique. Thus, it is important to notice, in the equations presented below, that although RsP is a set, it may have repeated elements, different from the traditional set theory. Thereby, in the following operations the resulting sets may contain some repeated rules.

Issue 1. Regarding the existing overlap among the rules in RsP and RsT, four metrics are proposed, which are described as follows:

M_{O-RsP} Measures the ratio of "old" rules in RsP, i.e., the ratio of rules in RsT found in RsP (Equation 1). A rule is considered "old" if it is in RsT, i.e., in the rule set obtained through the traditional process. Therefore, the higher

the value the better the metric, since the value indicates that there was no loss of knowledge during the process.

$$M_{O-RsP} = \frac{|RsT \cap RsP|}{|RsT|} . \tag{1}$$

M_{N-RsP} Measures the ratio of "new" rules in RsP, i.e., the ratio of rules in RsP not found in RsT (Equation 2). A rule is "new" if it isn't in RsT, i.e., in the rule set obtained through the traditional process. Although it is important that any knowledge be lost (metric M_{O-RsP}), it is expected that the ratio of "new" rules in RsP be greater than the ratio of "old" rules. Therefore, the higher the value the better the metric, since the value indicates the amount of knowledge, previously unknown, obtained during the process.

$$M_{N-RsP} = \frac{|RsP - RsT|}{|RsP|} . \tag{2}$$

$M_{R-O-RsP}$ Measures the ratio of "old" rules that repeat in RsP (Equation 3). It is considered that a rule should exist in only one of the clustering groups, since it has to be in a subdomain that expresses its own associations. Therefore, the lower the value the better the metric, since the value indicates that the knowledge, already known, is in subsets that express its own associations.

$$M_{R-O-RsP} = \frac{FindRepetitionRsP(RsT \cap RsP)}{|RsT \cap RsP|} , \tag{3}$$

FindRepetitionRsP: function that receives by parameter a set of non repeated rules and returns the number of rules in the set that repeat in RsP.

$M_{R-N-RsP}$ Measures the ratio of "new" rules that repeat in RsP (Equation 4). Idem to $M_{R-O-RsP}$. Therefore, as in $M_{R-O-RsP}$, the lower the value the better the metric, since the value indicates that the knowledge, previously unknown, is in subsets that express its own associations.

$$M_{R-N-RsP} = \frac{FindRepetition(RsP - RsT)}{|RsP - RsT|} , \tag{4}$$

FindRepetition: function that receives by parameter a set that may contain repeated rules and returns the number of rules in the set that repeat.

Issue 2. Regarding the existing overlap among the rules in RsP and RsT considering the interesting aspect of the knowledge, four metrics are proposed, which are described as follows:

$M_{N-I-RsP}$ Measures the ratio of "new" rules among the h-top interesting rules in RsP (Equation 5). Given a subset of h-top interesting rules, selected from RsP, it is expected that the ratio of "new" rules in this subset be as large as possible. The h-top rules are the h rules that contain the highest values regarding an objective measure, where h is a number to be chosen. Therefore, the higher the value the better the metric, since the value indicates that the cost of the process is minimized by the discovery of interesting knowledge, previously unknown, in RsP.

$$M_{N-I-RsP} = \frac{CountTopRules(h_{top\ of\ RsP}, RsP - RsT)}{|h_{top\ of\ RsP}|}\ , \tag{5}$$

CountTopRules: function that receives by parameter a set of h-top interesting rules and a set of rules and returns the number of rules that appears among the h-top.

$M_{O-I-N-RsP}$ Measures the ratio of "old" rules not in RsP among the h-top interesting rules in RsT (Equation 6). Given a subset of h-top interesting rules, selected from RsT, it is expected that all these rules are present in RsP. It is not desirable that the interesting patterns in RsT disappear in RsP, which would imply in the loss of relevant knowledge. Thus, this metric measures the ratio of "old" interesting rules not in RsP. The h-top rules are as described in $M_{N-I-RsP}$. Therefore, the lower the value the better the metric, since the value indicates that the interesting knowledge in RsT was not lost during the process.

$$M_{O-I-N-RsP} = \frac{CountTopRules(h_{top\ of\ RsT}, RsT - RsP)}{|h_{top\ of\ RsT}|}\ , \tag{6}$$

CountTopRules: idem Equation 5.

M_{C-I} Measures the ratio of common rules among the h-top interesting rules in RsP and the h-top interesting rules in RsT (Equation 7). Consider two subsets, S_1 and S_2, containing, respectively, the h-top interesting rules in RsP and the h-top interesting rules in RsT. This metric measures the existing intersection between these two subsets, which is expected to be as small as possible. Therefore, the lower the value the better the metric. The higher the intersection, the less relevant will be the process, since all the knowledge already known as interesting in RsT is also identified as interesting in RsP, not providing to the process any additional relevant information.

$$M_{C-I} = \frac{|h_{top\ of\ RsP} \cap h_{top\ of\ RsT}|}{h}\ , \tag{7}$$

h is the number to be chosen to realize the selection of the rules in both sets, i.e., RsP and RsT.

$M_{NC-I-RsP}$ Measures the ratio of groups in the clustering related to RsP that contains the h-top interesting rules in RsP (Equation 8). Therefore, the lower the value the better the metric. This means that just some of the groups would have to be explored by the user, which will contain the "new" relevant knowledge extracted during the process.

$$M_{NC-I-RsP} = \frac{FindGroups(h_{top\ of\ RsP})}{N}\ , \tag{8}$$

N: number of groups in the clustering; FindGroups: function that receives by parameter a set of h-top interesting rules, finds their groups and returns the number of distinct selected groups.

Issue 3. Regarding the process behavior related to the number of rules that are obtained in RsP, a unique metric is proposed, which is described as follows:

M_{NR-RsP} Measures the ratio of rules in RsP in relation to RsT (Equation 9). It is important to analyze the process in relation to the number of rules in RsP. It is not desirable to have a large increase in the volume of rules, because even if new patterns are discovered, it will be harder for the user to identify them. Therefore, the lower the value the better the metric, since the value indicates that although new patterns have been extracted, the number of extracted rules is not big enough to overload the user.

$$M_{NR-RsP} = \frac{|RsP|}{|RsT|} \,. \tag{9}$$

Relating the proposed metrics with the researchers found in the literature (Section 5), it can be observed that: (a) [1] is the only work that provides a similar analysis related to the metrics M_{O-RsP} and M_{N-RsP} in *Issue 1*; (b) none of them provide an analysis related to the aspects covered by *Issue 2*; [2,3,1] provide a similar analysis related to the metric M_{NR-RsP} in *Issue 3*. Thus, the necessity of a standardized assessment procedure becomes evident (more details in Section 5). Finally, it is important to say that we believe that these nine metrics cover, adequately, the three presented issues. However, as other issues arise, new metrics can be added to this assessment procedure.

3 Experiments

Some experiments were carried out in order to present how the metrics can be used. For that, two contexts were defined. Suppose a user decides to apply clustering in the pre-processing step. First of all, he has to find out the most suitable methodology to be used in his problem. After that, he has to check if the selected methodology was good enough for the problem, considering that different interests may be important for his decision. Thus, two different situations were regarded: (i) identify among some organizations the most suitable; (ii) analyze the process itself. An organization is obtained by the application of a clustering algorithm combined with a similarity measure. Therefore, the metrics provide the support to evaluate each situation under the discussed issues: while in (i) the data is initially clustered through some organizations in order to identify the organization that obtains a good association set, in (ii) the usefulness of the process itself is analyzed. Four data sets and four organizations were selected to be used in the experiments.

The four data sets were Adult (48842;115), Income (6876;50), Groceries (9835;169) and Sup (1716;1939). The numbers in parenthesis indicate, respectively, the number of transactions and the number of distinct items in each data set. The first three are available in the R Project for Statistical Computing through the package "arules"[1]. The last one was donated by a supermarket located in São Carlos city, Brazil. All the transactions in Adult and Income contain the same number of items (named here as standardized data sets (S-DS)), different from Groceries and Sup (named here as non-standardized data sets

[1] `http://cran.r-project.org/web/packages/arules/index.html`.

(NS-DS)), where each transaction contains a distinct number of items. Thus, the experiments considered different data types.

The four organizations were obtained by the combination of the algorithms and similarity measures presented in Table 1. Each combination gives an organization, i.e., a different way to analyze the process. Although it is necessary to set k, the number of groups to be generated, in order to obtain an organization, this value was used to analyze the organizations on different views. Despite the existence of algorithms designed for transactions, such as ROCK, the choices of the algorithms were made based on works that cluster the rules in the post-processing phase aiming a *posteriori* comparison. The similarity measures were chosen considering the works described in Section 5 – only the similarities among transactions were selected.

As described before, the rules are extracted within each group after clustering the data. The values of the minimum support (min-sup) and minimum confidence (min-conf) have to be set in order to extract a set of association rules. To automate the specification of the min-sup in each group, the following procedure was adopted: (i) find the 1-itemsets of the group with their supports, (ii) compute the average of these supports, (iii) use this average support as the min-sup of the group. Regarding min-conf, the following values were used for each data set: Adult 50%; Income 50%; Groceries 10%; Sup 100%. Thus, the same min-conf value was applied in all the groups of a given data set. These values were chosen experimentally. Although it is known that min-sup and min-conf impacts on the set of rules that are obtained, it was assumed that the focus was on the use of the metrics and, so, that the values were adequate to the proposed problem. Finally, the rules were extracted with an Apriori implementation developed by Christian Borgelt[2] with a minimum of two items and a maximum of five items per rule.

Considering the four organizations, the RsP sets were obtained. However, once almost all the metrics are based on the rules obtained through the traditional process, the four data sets were also processed to obtain the RsT sets. For that, the min-sup was set automatically, as described before. Regarding min-conf, the same values used in RsP were considered, i.e., Adult 50%, Income 50%, Groceries 10% and Sup 100%. Furthermore, as some of the metrics are based on the h-top interesting rules of a given rule set, an objective measure should be selected. Instead of choosing a specific measure, the average rating obtained through 18 objective measures (see Table 1) was considered as follows: (i) the value of 18 measures was computed for each rule; (ii) each rule received 18 ID's, each ID corresponding to the rule position in one of the ranks related to a measure; (iii) the average was then calculated based on the rank positions (ID's). Thus, the h-top rules were selected considering the best average ratings. h, also a number to be set, was defined, in all the sets (RsT and RsP), to assume 0.5% of the total of rules in RsT – as seen in Section 4, always the smallest set. Therefore, each rule set contains its own values that are proportional in all of them. Table 1 summarizes the configurations used to apply the proposed metrics.

[2] `http://www.borgelt.net/apriori.html`.

Table 1. Configurations used to apply the proposed metrics

Data sets	Adult; Income; Groceries; Sup
Algorithms	PAM; Ward [algorithms details in [5]]
Similarity measures	Agg; Denza
k	5 to 25, steps of 5
h	0.5% of the total of rules in RsT
Objective measures [measures details in [6]]	Added Value, Certainty Factor, Collective Strength, Confidence, Conviction, IS, ϕ-coefficient, Gini Index, J-Measure, Kappa, Klosgen, λ, Laplace, Lift, Mutual Information (asymmetric), Novelty, Support, Odds Ratio

4 Results and Discussion

Considering the configurations presented in Table 1 and the RsT sets above described, the experiments were carried out and the values of each metric obtained. Regarding the first proposed situation, i.e., identify among some organizations the most suitable (Section 3), an analysis based on the average of each metric, considering the different data types, apart from the data set, was carried out. Table 2 presents the results. Thus, in this case, the metrics will help the users to find out a suitable methodology for their problems. In order to aid the comparison of the results, all the metrics that present better results when their values are the smallest ($M_{R-O-RsP}$, for example) were processed to store the complement of the information. Therefore, all the metric, with exception to M_{NR-RsP}, have the same interpretation: the higher the value the better the performance. Furthermore, all the metrics can be seen in terms of percentage if multiplied by 100 (ex.: 0.858*100 = 85.8%).

Table 2. Average of the proposed metrics in the considered organizations

Data type	Algorithm	Measure	M_{O-RsP}	M_{N-RsP}	$M_{R-O-RsP}$	$M_{R-N-RsP}$	$M_{N-I-RsP}$	$M_{O-I-N-RsP}$	M_{C-I}	$M_{NC-I-RsP}$	M_{NR-RsP}
S-DS	PAM	Agg	0.858▲	0.906	0.239△	0.881△	0.716	0.891▲	0.838	0.596△	125.370▲
		Denza	0.730	0.936△	0.235	0.878	0.874▲	0.583	0.912△	0.563	160.534
	Ward	Agg	0.718	0.923	0.213△	0.871	0.920△	0.466	0.967△	0.503	129.681▲
		Denza	0.722△	0.928△	0.209	0.877△	0.870	0.509△	0.919	0.533△	133.994
NS-DS	PAM	Agg	0.880	0.901△	0.709▲	0.974△	0.750△	1.000△	1.000△	0.909▲	269.830▲
		Denza	0.924△	0.885	0.510	0.940	0.745	0.986	0.946	0.785	431.619
	Ward	Agg	0.976△	0.245	0.947▲	0.999△	0.211	0.997△	0.297	0.828	1.652▲
		Denza	0.973	0.693▲	0.755	0.974	0.604▲	0.997△	0.684▲	0.867△	221.917

Each average in Table 2 was obtained from the results of the experiments related to the presented configuration. The value 0.858 in M_{O-RsP} at S-DS:-PAM:Agg, for example, was obtained by the average of the values in M_{O-RsP} at Adult:PAM:Agg and Income:PAM:Agg over the values of k. The highest averages are marked with △ in each algorithm regarding each metric. The only exception is M_{NR-RsP}, where the lowest averages are highlighted. For the S-DS:PAM configuration, for example, the best average for M_{O-RsP} is the one related to Agg (0.858). However, since the averages are, in general, near, a marking based on the difference among the averages was also considered. The values marked with ▲ indicate that the difference between the averages of a given metric are

above 0.1 ($dif. \geq 0.1$). For the S-DS:PAM configuration, for example, the best average for M_{O-RsP} is the one related to Agg (0.858), presenting a difference of 0.128 in relation to Denza (0.858-0.730). Thereby, it is possible to visualize, for each configuration, the most suitable similarity measure. It is important to mention that the results are deterministic and, therefore, no statistical test was done to check if there is a significant difference among the averages. It can be noticed that:

S-DS:PAM. The most suitable measure for this configuration is Agg, since it presents better results in 6 of the 9 metrics in relation to Denza. Furthermore, in 3 of the 6 metrics Agg exhibits a difference above 0.1 in relation to Denza. In these cases, it can be noticed that the values in Agg are more representative than the values in Denza – observe, for example, that while Agg in $M_{O-I-N-RsP}$ presents a performance above 89%, Denza presents a performance below 59%.

S-DS:Ward. Although Denza presents a better performance in relation to Agg in 5 of the 9 metrics, Agg seems to be the most suitable measure for this configuration even presenting a better performance in 4 of the 9 metrics. This occurs because while Agg exhibits a difference above 0.1 in relation to Denza in 1 of the 4 metrics, Denza doesn't present any difference in any of the metrics. Although the difference occurs in only one of the metrics, the metric is important, since it measures how much the exploration space increases in relation to RsT.

NS-DS:PAM. The most suitable measure for this configuration is Agg, since it presents better results in 8 of the 9 metrics in relation to Denza. Furthermore, in 3 of the 8 metrics Agg exhibits a difference above 0.1 in relation to Denza. In these cases, it can be noticed that the values in Agg are more representative than the values in Denza – observe, for example, that while Agg in $M_{R-O-RsP}$ presents a performance above 70%, Denza presents a performance below 52%.

NS-DS:Ward. Both measures present a good performance in 4 of the 9 metrics, excluding the tie occurred in $M_{O-I-N-RsP}$. However, in 3 of the 4 metrics Denza exhibits a difference above 0.1 in relation to Agg. In these cases, it can be noticed that the values in Denza are more representative than the values in Agg – observe, for example, that while Denza in M_{N-RsP} presents a performance above 69%, Agg presents a performance below 25%. Therefore, the most suitable measure for this configuration is Denza.

Considering the exposed arguments, it can be noticed that:

S-DS. Comparing the results of PAM:Agg and Ward:Agg, PAM presents better results in 6 of the 9 metrics in relation to Ward (M_{O-RsP}, $M_{R-O-RsP}$, $M_{R-N-RsP}$, $M_{O-I-N-RsP}$, $M_{NC-I-RsP}$, M_{NR-RsP}) and a difference above 0.1 in 3 of the 6 metrics (M_{O-RsP}, $M_{O-I-N-RsP}$, M_{NR-RsP}). Therefore, the most suitable organization, to this type of data, according to the metrics, is PAM:Agg.

NS-DS. Comparing the results of PAM:Agg and Ward:Denza, PAM presents better results in 5 of the 9 metrics in relation to Ward (M_{N-RsP}, $M_{N-I-RsP}$, $M_{O-I-N-RsP}$, M_{C-I}, $M_{NC-I-RsP}$), excluding the tie occurred in $M_{R-N-RsP}$, and a difference above 0.1 in 3 of the 5 metrics (M_{N-RsP}, $M_{N-I-RsP}$, M_{C-I}).

Therefore, the most suitable organization, to this type of data, according to the metrics, is also PAM:Agg.

As observed, the most suitable organization according to the metrics, regarding the presented configurations (Table 1), apart from the data type used, is PAM:Agg. In other words, the user will obtain better results, i.e., reasonable rule set, if he initially clusters the data through PAM:Agg. However, in other domains, different aspects can be of interesting, providing the user a flexible way to solve his issues. Thus, in this first situation, the metrics provide criteria to carry out comparisons, helping the user to select the most suitable methodology for his problem.

From that point, supposing that PAM:Agg is a suitable solution for the user's problem, it is possible to analyze the process itself, i.e., to check if good results are really obtained. Observe that different interests may be important for his decision. Thus, the metrics provide criteria not only to analyze the process, but also to identify its positive and negative aspects, helping the user to reach a conclusion. To discuss this second situation, Table 3 presents the values of the metrics, in the selected organization, in each one of the data types. These values are the ones presented at S-DS:PAM:Agg and NS-DS:PAM:Agg in Table 2, but in their original scales, since the smaller scales (↓) were previously converted – the larger scales (↑) remain the same. The scale, in each metric, is found between "[]". It can be noticed that:

M_{O-RsP} Little knowledge is lost during the process, around 15%, since more than 85% of the rules in RsT are found in RsP. Thus, a positive aspect of the process is identified.

M_{N-RsP} Almost all the rules in RsP are "new", around 90%, indicating the discovery of a great amount of knowledge previously unknown. Thus, a positive aspect of the process is identified.

$M_{R-O-RsP}$ The repetition of "old" rules in RsP is high in both data types, around 30% at NS-DS and 77% at S-DS. Thus, a negative aspect of the process is identified.

$M_{R-N-RsP}$ The repetition of "new" rules in RsP is low, around 12%, indicating that the knowledge, previously unknown, is in subdomains that express their own associations. Thus, a positive aspect of the process is identified.

$M_{N-I-RsP}$ A great amount of the h-top interesting rules in RsP are "new", around 71%, indicating that the cost of the process is minimized by the discovery of interesting knowledge, previously unknown, in RsP. Thus, a positive aspect of the process is identified.

$M_{O-I-N-RsP}$ The loss of "old" and interesting knowledge is low, around 11%, since a great amount of the h-top interesting rules in RsT are found in RsP, around 89% (100%-11%). Thus, a positive aspect of the process is identified.

M_{C-I} The intersection between the h-top interesting rules in RsP and the h-top interesting rules in RsT is low, around 17%, indicating that few of the knowledge already known as interesting in RsT is found in RsP. Thus, a positive aspect of the process is identified.

$M_{NC-I-RsP}$ The number of groups that contain the h-top interesting rules in RsP at NS-DS is low, around 10%, which doesn't occur at S-DS, that is around 41%. Thus, a negative aspect of the process is identified.

M_{NR-RsP} The number of rules in RsP is far greater in relation to RsT, overloading the user with an excessive number of rules. Thus, a negative aspect of the process is identified.

Table 3. Average of the proposed metrics in the PAM:Agg organization

Data type	M_{O-RsP} [↑]	M_{N-RsP} [↑]	$M_{R-O-RsP}$ [↓]	$M_{R-N-RsP}$ [↓]	$M_{N-I-RsP}$ [↑]	$M_{O-I-N-RsP}$ [↓]	M_{C-I} [↓]	$M_{NC-I-RsP}$ [↓]	M_{NR-RsP} [↓]
S-DS	0.858	0.906	0.761	0.119	0.716	0.109	0.162	0.404	125.370
NS-DS	0.880	0.901	0.291	0.026	0.750	0.000	0.000	0.091	269.830

Summarizing, it can be observed that: (a) a great amount of interesting knowledge, previously unknown, is discovered, which are in subdomains that express their own associations; (b) little interesting knowledge is lost, which are not in subdomains that express their own associations; (c) since the number of rules is high and the interesting knowledge, previously unknown, is spread over the clustering groups, the user exploration can be hampered. Therefore, considering the positive and negative aspects of the process, the user can analyze the results, according to his interests, and conclude if good results were reached. It is relevant to say that the importance of each percentage depends on the user's needs, on the data sets, etc., and, therefore, has to be, in fact, validate by them. Regarding the presented context, it can be said that the process obtains reasonable results, since 6 of the 9 aspects were considered positives. However, if the weight of the 3 negative metrics is more important to the user, he can discard the results. Moreover, he can try to improve the process to obtain better results in these metrics. Thus, in this second situation, the metrics provide criteria to analyze the process based on different interests, identifying its positive and negative aspects, helping the user to reach a conclusion.

5 Related Works

There are many researches that initially cluster data aiming to discover and facilitate the search for the interesting pattern of the domain. Some of these works are described below.

[2] propose to split the data set items into groups in order to extract the rules. The authors evaluate many hierarchical algorithms combined with many similarity measures. Nevertheless, it is not understandable how the rules are obtained within the groups, since it is necessary to have a set of transactions and not a set of items. This means that it is not clear how the transactions are distributed over the groups. Among the similarity measures used by them, we emphasize Jaccard due to its use by the measure described below (Agg). The Jaccard between two items i_1 and i_2, expressed by P-J(i_1,i_2)=$\frac{|\{t \text{ covered by } i_1\} \cap \{t \text{ covered by } i_2\}|}{|\{t \text{ covered by } i_1\} \cup \{t \text{ covered by } i_2\}|}$, is the

ratio between the transactions t the items cover simultaneously and the total of transactions the items cover. An item covers a transaction t if the item is in t.

[3] propose an algorithm, named CLASD, to split the data set transactions aiming to discover associations on small segments (subsets) of the data. To cluster the transactions, a similarity measure proposed by them, named Agg, expressed by $\text{Agg}(t_1,t_2)=\frac{\sum_{p=1}^{m}\sum_{q=1}^{n} Af(i_p,j_q)}{m*n}$, is used. Thus, the similarity between two transactions t_1 and t_2 is computed by the affinity (Af) average among the transaction items, being Af equivalent to the measure P-J. Therefore, after computing P-J among the m items in t_1 and the n items in t_2, the average among them is obtained.

[1] propose an algorithm, named *Apriori Inverse*, to cluster the transactions and then extract a rare association rule set. To cluster the transactions, their algorithm initially finds k seeds (centroids), where k indicates the number of frequent itemsets that match some conditions. Each seed forms a group. After the seed generation, each transaction t is allocated to one of the groups based on the number of common items that occur between the transaction (t) and the group centroid.

There are other researches concerned with the clustering of transactions that, although not related to the extraction of association rules, could be used. In [4], for example, the authors propose an approach to identify, a priori, the potentially interesting items to appear in the antecedents and in the consequents of the association rules without extracting them. The approach is divided in two steps: the clustering of the transactions and the selection of the interesting items. To do the clustering the authors propose the use of incremental K-means with a similarity measure obtained through a Jaccard between transactions, expressed by $\text{Denza}(t_1,t_2)=\frac{|\{items\ in\ t_1\}\cap\{items\ in\ t_2\}|}{|\{items\ in\ t_1\}\cup\{items\ in\ t_2\}|}$. Therefore, the similarity between two transactions t_1 and t_2 is computed considering the items the transactions share.

Among the papers above described, little has been done to analyze the performance of the methodologies, allowing to identify their positive and negative aspects, or even to compare the results among them. In general, the researchers compare the number of rules and/or itemsets that are obtained from unpartitioned data and clustered data to expose the usefulness of the methodologies. This strategy can be found in [2,3,1] and is related to "Issue 3" of Section 1. However, [2] also analyze the process considering the complexity of the rules that are obtained – the greater the number of items that compose a rule the higher its complexity. [3,1] discuss over some rules found through clustering to show that the process provides the discovery of interesting patterns, but the analysis of the process is subjective. [3] also consider the execution time. Finally, [1] is the only work that allows a better analysis considering the existing overlap between the rules obtained from unpartitioned data and clustered data. This strategy is related to "Issue 1" of Section 1. Based on the presented arguments, as mentioned before, the necessity of a standardized assessment procedure becomes evident.

6 Conclusion

In this paper, nine metrics were proposed to support the evaluation of methodologies that use clustering in the pre-processing step to aid the discovery of associative interesting patterns. The metrics were developed to answer three main issues related to the presented context. Some experiments were carried out in order to present how the metrics can be used. For that, two different situations were regarded: (i) identify among some organizations the most suitable; (ii) analyze the process itself. Through the experiments, it could be noticed that the metrics provide criteria to: (a) analyze the methodologies, (b) identify their positive and negative aspects, (c) carry out comparisons among them and, therefore, (d) help the users to select the most suitable solution for their problems. Thus, based on the discussions, the usefulness and the importance of the metrics were demonstrated.

As a future work, to complement the results, an empirical study with human subjects will be done to verify if configurations with high metric values indeed produces rules that are more useful to end users. With this new analysis, we believe that a better understanding of the presented context will be reached and its importance highlighted. A new methodology that tries to optimize all these criteria, through an optimization technique, can be an interesting proposal.

Acknowledgments. We wish to thank Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (processes numbers: 2010/07879-0 and 2011/19850-9) for the financial support.

References

1. Koh, Y.S., Pears, R.: Rare association rule mining via transaction clustering. In: 7th Australasian Data Mining Conference. CRPIT, vol. 87, pp. 87–94 (2008)
2. Plasse, M., Niang, N., Saporta, G., Villeminot, A., Leblond, L.: Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Computational Statistics & Data Analysis 52(1), 596–613 (2007)
3. Aggarwal, C.C., Procopiuc, C., Yu, P.S.: Finding localized associations in market basket data. IEEE Transactions on Knowledge and Data Engineering 14(1), 51–62 (2002)
4. D'Enza, A.I., Palumbo, F., Greenacre, M.: Exploratory data analysis leading towards the most interesting binary association rules. In: 11th Symposium on Applied Stochastic Models and Data Analysis, pp. 256–265 (2005)
5. Xu, R., Wunsch, D.: Clustering. IEEE Press Series on Computational Intelligence. Wiley (2008)
6. Tan, P.-N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Information Systems 29(4), 293–313 (2004)