# New Measure for Evaluating Association Rules

**Conference Paper** · August 2014

DOI: 10.1109/SOCPAR.2014.7008041

Youcef Djenouri
SINTEF As
**141** PUBLICATIONS **2,120** CITATIONS

SEE PROFILE

Youcef Gheraibia
York St John University
**26** PUBLICATIONS **269** CITATIONS

SEE PROFILE

Malika Mehdi
University of Science and Technology Houari Boumediene
**19** PUBLICATIONS **163** CITATIONS

SEE PROFILE

Ahcene Bendjoudi
Research Center on Scientific and Technical Information
**47** PUBLICATIONS **411** CITATIONS

SEE PROFILE

# An Efficient Measure for Evaluating Association Rules

Youcef Djenouri, Youcef Gheraibia, Malika Mehdi, Ahcene Bendjoudi, Nadia Nouali-Taboudjemat
*CERIST Center Research, Algiers, Algeria*
*Dept Math and Computer science, Univ Souk Ahras, Algeria*
*USTHB LSI, Algiers, Algeria*
*ydjenouri@cerist.dz, youcef.gheraibia@cu-soukahras.dz, malika.mehdi@gmail.com, abendjoudi@cerist.dz, nnouali@cerist.dz*

***Abstract*—Association rules mining (ARM) has attracted a lot of attention in the last decade. It aims to extract a set of relevant rules from a given database. In order to evaluate the quality of the resulting rules, existing measures, such as support and confidence, allow to evaluate the resulted rules of ARM process separately, missing the different dependencies between the rules. This paper addresses the problem of evaluating rules by taking into account two aspects: (1) The accuracy of the returned rules on the input data and (2) the distance between the returned rules. The rules set that covers the maximum of rules space is considered. To analyze the behavior of the proposed measure, it has been tested on two recent ARM algorithms BSO-ARM and HBSO-TS.**

**Keywords**:Association rules mining, Rules Quality, Evaluation of Rules

## I. Introduction

Association Rules Mining (ARM) is one of the most important and well studied techniques of data mining tasks [1]. It aims to extracting frequent patterns, associations or causal structures among sets of items from a given transactional database. Formally, the association rule problem is formulated as follows: let $T$ be a set of transactions $\{t_1, t_2, \dots, t_m\}$ representing a transactional database, and $I$ be a set of $m$ different items or attributes $\{i_1, i_2, \dots, i_m\}$, an association rule is an application $X \to Y$ where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$. The itemset $X$ is called antecedent while the itemset $Y$ is called consequent and the rule $X \to Y$ means $X$ implies $Y$.

There exist different measures in the literature to evaluate the resulted rules of an ARM algorithm [3], [1], [16]. Nevertheless, all these measures compute the frequency of the rules from the given data. This aspect of the classical evaluation measures does not reflect enough the quality of the rules. Indeed, there could be some interesting rules with low frequency and evident rules with high frequency [17]. Also, the resulted rules could be too close to each other. In fact, they could describe the same transactions or the same items.

This paper deals with this problem by taking into account the concept of cover rules. The cover rules are the rules that cover the maximum of rules space. The most distant rules are the most cover rules. In addition, two formulas for computing the distance between rules are proposed. The first formula is called ID(Items-based Distance). Using ID, the cover rules are the rules that contain the maximum of items. The second formula is called DRD (Data Rows-based Distance). Using this one, the cover rules are the rules describing the maximum of transactions. Using the concepts of distance and cover rules, in this paper we propose a new evaluation measure based on both frequency and distances. The experiments reported in this paper show the effectiveness of ID with high number of transactions and the effectiveness of DRD with high number of items. In addition, we analyzed the behavior of the two distances on two recent ARM algorithms (BSO-ARM and HBSO-TS). The results revealed the convergence of the two algorithms in terms of the resulted rules set using the both distances and on the most input datasets.

The rest of this paper is organized as follows. The next section relates about association rules mining algorithms. In section II-B, the most used measures of ARM problem are presented. In section III, the new measure of ARM problem is presented. Section IV reports on experimental results of our measure and the behavior of the two proposed distances on ten datasets with different sizes using two recent algorithms. Finally, Section V concludes this paper by some remarks and some perspectives for a future work.

## II. Related Works

This section gives a brief overview about two topics related to this paper: (1) ARM algorithms including well known exact algorithms, such as Apriori, and some new proposed methods based on meta-heuristics, (2) the most used statistical measures used to evaluate relevant rules generated by ARM algorithms.

### A. ARM algorithms

Many algorithms for generating association rules have been proposed in literature. Some well known exact algorithms are AIS [2], Apriori [4], Eclat [5] and FP-Growth [6]. Apriori is the most used algorithm for association rules mining. It is based on breath first search

978-1-4799-5934-1/14/$31.00 ©2014 IEEE

strategy to count the supports of itemsets and uses a candidate generation function to exploit the downward closure property of support. In [4], the authors proposed a new method based on Apriori algorithm for generating all positive and negative association rules, called NRGA which generate all hidden association rules.
However, because of the growth of transactional databases, exact methods have become inefficient on large databases. Consequently, approximate methods, especially meta-heuristics, are more and more used for ARM. Indeed, such methods allow to deal with large databases in less time compared to the traditional ARM algorithms.

Different meta-heuristics have been applied for ARM problem. Yan et al. [7] developed a genetic algorithm called ARMGA, for finding association rules. The main drawback with this algorithm is the generation of non admissible solutions and hence erroneous rules that may not respect Minimum Support and Minimum Confidence constraints. In [8], another GA is proposed for Mining Association rules. By using an adaptive mutation rate, this algorithm provides an important population variation. An improved version of genetic algorithm is used to consider negative association rules, the aim of this algorithm is to find all possible optimized rules from given data set [15]. In [9], an Adaptive Genetic Algorithm called AGA is developed for computing ARM. In [10], the authors proposed a new Bees Swarm Optimization algorithm for ARM (BSO-ARM) which avoids generating false rules and solves the admissibility problem by defining an efficient representation of the solution and a strong fitness function. It uses the bees behavior on the exploration of the rules space. In [11], a hybrid method combining a BSO algorithm and a Tabu Search algorithm (HBSO-TS) is proposed for ARM. In HBSO-TS, the main process is done using BSO-ARM. However, the exploration of the region of each bee is performed by the tabu search. The results revealed the efficiency of HBSO-TS compared to BSO-ARM.
Nonetheless, the optimization methods applied to ARM give only one part of the relevant rules and not all possible rules . Consequently, the existing formulas do not measure really the quality of the resulted rules.
For this, our purpose is to find a formula that measures efficiently the rules obtained by the ARM process.

### B. Statistical Measures for Association Rule Mining Problem

There are different measures to evaluate the association rules obtained by ARM process.

Two basic formulas are commonly used for measuring usefulness of association rules, namely the support of a rule and a confidence of a rule. The support of an itemset $I' \subseteq I$ is the number of transactions containing $I'$. The support of a rule $X \rightarrow Y$ is the support of $X \cup Y$ and the confidence of a rule is:

$$\frac{support(X \cup Y)}{support(X)}$$

Confidence is a measure of strength of the association rules. An association rule $X \rightarrow Y$ with a confidence of $80\%$ means that $80\%$ of the transactions that contain $X$ also contain $Y$ together. So, the aim is to extract from a given database, all interesting rules, that is rules with support $\geq$ MinSup and confidence $\geq$ MinConf[3] where MinSup and MinConf are two thresholds predefined by users.
Other measures have been proposed in the literature like Lift, Leverage and Coverage[1] as follows:
$Lift(X \rightarrow Y) = \frac{confidence(X \rightarrow Y)}{support(Y)}$.
$Leverage(X \rightarrow Y) = support(X \rightarrow Y) - (support(X) \times support(Y))$.
$Coverage(X \rightarrow Y) = support(X)$.
All these measures are based only on the frequency of the rules in the transactional database. However, it exists evident rules with high frequency and rare rules with low frequency. In addition, it is possible to find a set of rules with high frequencies but which describe the same transactions. When dealing with large transactional databases, the aim is to find a set of rules that describe the maximum of transactions. It means to find rules that cover the maximum of rules space. The classical measures presented in this section do not take these problems into account. In the next section, we propose a new measure which allows to evaluate the association rules by considering all these issues.

## III. The proposed Measure

Assume $\omega = \{r_1, r_2, ...r_p\}$ be the rules space containing $p$ association rules. Association rules mining process aims to find the set of pertinent rules covering the maximum of rules in the rules space. The quality of the rules is measured using the statistical measures described in section II-B. Whereas, the set of rules is measured using the notion of the similarity between the obtained rules. The farther are the rules from each other, the better is the coverage of the set on the rules space $\omega$.

Let $R = \{r_1, r_2, ...r_k\}$ be a subset of $\omega$ containing the $k$ rules obtained by ARM process.
The quality of the set $R$ can be computed as follows

$$Evaluate_{max}(R) = Statistical_{max}(R) + Cover_{max}(R). \quad (1)$$

where

$$Statstical_{max}(R) = \frac{\sum_{r_i \in R} Confidence(r_i)}{k}. \quad (2)$$

$$Cover_{max}(R) = \frac{\sum_{r_i \in R} \sum_{r_j \in R} Distance(r_i, r_j)}{k}. \quad (3)$$

According to Equation 1, the measure aims to maximize both statistical measure and the coverage of the rules space. The function Cover calculates the distances of all resulted rules. The aim is to maximize this function in order to cover the maximum of rules space. This function depends on the definition of the distance between rules. There are different strategies to measure the distance between rules. Each strategy gives a particular state of the rules space. In the following, we propose two strategies to measure the distance between the rules and we analyze the effects of these two strategies on the rules space.

- Items-based Distance (ID) The distance ID defines the similarities between two rules according to the items contained in these rules. The rules are similar when they share many items. They are dissimilar where they do not share any item. Formally, the ID formula is given as follows:

$$ID(r_1, r_2) = |items(r_1) \cup items(r_2)| - |items(r_1) \cap items(r_2)| \quad (4)$$

  *Example 1:* Let us consider five different items{A, B, C, D, E} and the two rules
  $r_1 : A, B, C \Rightarrow D$
  $r_2 : C \Rightarrow E$
  First, we define the set of items of the two rules:
  items($r_1$)={A, B, C, D}
  items($r_2$)={C, E}
  Then, we determine the union of the two sets of items and the common items of these two rules as
  items($r_1$)∪ items($r_2$)={A, B, C, D, E}.
  items($r_1$)∩ items($r_2$)={C}.
  So, The distance ID between $r_1$ and $r_2$ is
  ID($r_1$, $r_2$)= $5 - 1 = 4$.
- Data Rows-based Distance (DRD) The distance DRD defines the similarities between two rules according to the number of common transactions between these rules. The rules are close when they share many transactions, in other terms, they verify together many transactions. They are dissimilar where they do not share any transaction. Formally, the DRD formula is given as follows:

$$DRD(r_1, r_2) = m - |T(r_1) \cap T(r_2)| \quad (5)$$

  Where
  $T(r_i)$ is the set of transactions verified by $r_i$
  *Example 2:* Let us consider five different items{A, B, C, D, E} and the five following transactions
  $t_1$: A, B, C, D
  $t_2$: A, C
  $t_3$: A, B, D
  $t_4$: C, B, D
  $t_5$: C, D, E
  After mining process, we assume the two obtained rules as follows:
  $r_1 : A, B, C \Rightarrow D$
  $r_2 : C \Rightarrow D$
  First, we define the set of transactions verified by each rule:
  T($r_1$)={$t_1$}
  T($r_2$)={$t_1$, $t_4$, $t_5$}
  Then, we determine the common transactions of the two rules as
  T($r_1$)∩ T($r_2$)={$t_1$}.
  So, The distance DRD between $r_1$ and $r_2$ is
  DRD($r_1$, $r_2$)= $5 - 1 = 4$.

Using the distance ID, the ARM algorithm aims to find the frequent rules which cover the maximum of items. However, using the distance DRD, the ARM algorithm aims to find the frequent rules which cover the maximum of transactions.

| Data Set Name | Trans Size | Item Size | Dataset type |
|---|---|---|---|
| Bolts | 40 | 8 | small |
| Sleep | 56 | 8 | small |
| Pollution | 60 | 16 | small |
| Basket ball | 96 | 5 | small |
| IBM.Q.St | 1000 | 40 | small |
| Quake | 2178 | 4 | small |
| Chess | 3196 | 75 | small |
| Mushroom | 8124 | 119 | small |
| BMS1 | 59602 | 497 | Large |
| Connect | 100000 | 999 | Large |

Table I
BENCHMARK DESCRIPTION

## IV. RESULTS AND DISCUSSION

### A. Datasets description

For the convenience of comparison, we conducted our experiments on real datasets to analyze the behavior of the proposed measure. The used dataset are the well-known benchmarks employed with other algorithms in experimentation, these datasets are divided on two categories, small datasets with few number of items and transactions and large datasets with high number of transactions. The datasets are derived from Frequent and Mining Dataset Repository [13] and Bilkent University Function Approximation Repository [12]. Table I presents the description of different datasets used later in the experimentation.

### B. Experimentation

In order to evaluate the performance of our new measure, several comparing done with the different datasets. First, we analyze the performances of ID and DRD distances. Figure 1 shows the execution times of DRD and ID distances by fixing the number of items to 1000 and varying the number of transactions from 5000 to 100000 transactions.

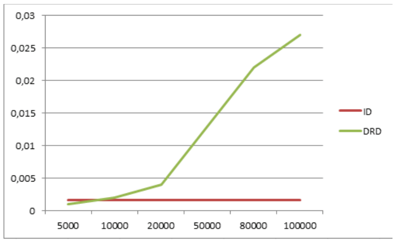

Figure 1. Execution time with different number of transactions (Sec)

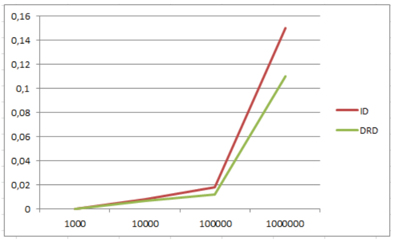

Figure 2. Execution time with different number of items (Sec)

According to this figure, we remark that DRD outperforms ID when few number of transactions is used. However, with high number of transactions, DRD consumes a considerable time compared to ID. In fact, the number of transactions affects the performances of DRD. This aspect is ignored in ID distance. Figure 2 presents the execution time of the two distances. We fix the number of transactions to 1000 and we vary the number of items from 1000 to 1 million. DRD outperforms ID in all cases. Indeed, the number of items affects considerably the ID performances. After this first experiment, we can say that the number of transactions affects DRD and the number of items affects ID. We conclude that ID will be used if we dispose low number of items, and DRD will be used if we have a low number of transactions. Now, the behavior of the proposed measure is analyzed on two recent bio-inspired algorithms for association rules mining that we proposed in two previous works. The first algorithm is BSO-ARM [10]. The search of relevant rules is performed using the bees swarm optimization. The second algorithm is HBSO-TS [11]. The whole process is done using BSO-ARM while the search of each bee is computed by the tabu search method. The experiments reported in [11] and [10] showed that the two algorithms outperform the state-of-the-art ARM algorithms using small and medium datasets.

In this paper, the proposed measure described in Section III is used as the fitness evaluation function on the two algorithms. Table II presents the results of the BSO-ARM and HBSO-TS algorithms in terms of solution quality by applying the proposed measures developped on Section III. According Table II, we observe a similarity between the results obtained by the two algorithms on all datasets except on the bolts and pollution datasets. On Bolts, BSO-ARM outperforms HBSO-TS, whereas on pollution base, HBSO-TS outperforms BSO-ARM using ID distance and BSO-ARM outperforms HBSO-TS using DRD distance.

| Data Set Name | BSO-ARM ID | HBSO-TS ID | BSO-ARM DRD | HBSO-TS DRD |
|---|---|---|---|---|
| Bolts | 8.2 | 8 | 15.2 | 15 |
| Sleep | 3.2 | 8 | 15 | 15 |
| Pollution | 4.8 | 8 | 20 | 19.80 |
| Basket ball | 8.2 | 8.2 | 20.2 | 20.2 |
| IBM.Q.St | 8 | 8 | 30 | 30 |
| Quake | 7.4 | 7.4 | 35.2 | 35.2 |
| Chess | 6.4 | 8 | 40 | 40 |
| Mushroom | 8 | 8 | 45 | 45 |
| BMS1 | 6.4 | 6.4 | 50 | 50 |
| Connect | 8 | 8 | 55 | 55 |

Table II
SOLUTION QUALITY OF BSO-ARM AND HBSO-TS ALGORITHMS USING THE PROPOSED MEASURE

## V. CONCLUSION

In this paper, we propose a new measure to evaluate relevant rules in ARM algorithms in an efficient way. Our measure takes into account higher quality rules that cover the maximum of rules space. Two measures for computing the distance between two rules are proposed. The first distance (ID) calculates the distance between rules according to the number of items shared by these rules. Using ID in the evaluation, the ARM algorithm allows to find rules covering the maximum of items. The second distance (DRD) determines the distance between rules according to the common number of transactions verified by each rule. In this case, the ARM algorithm gives rules describing the maximum of transactions.

The experiments revealed that ID is faster than DRD with high number of transactions and DRD is faster than ID with high number of items. The results also showed the convergence of the returned rules set of the BSO-ARM and HBSO-TS using the both distances.

As a perspective, we plan to develop a new algorithm for association rules mining problem based on the proposed evaluation measure.

## REFERENCES

[1] Han,J., Kamber, J. and Pei, M. (2011) Data Mining: Concepts and Techniques. Elsevier 3rd edition, pp 1-50

[2] Agrawal, R., Imielinski, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ACM SIGMOD Record (Vol. 22, No. 2, pp. 207-216). ACM.

[3] Agrawal, R. and Shafer, J. (1996) Parallel mining of associations rules. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL 8, NO 6, pp962-969

[4] Agrawal, R. and Ramakrishan,S.: *Fast algorithms for association rules in large databases*, Proc of the 20th International Conference on very large Data bases -VLDB), Santiago, Chile, PP 487-499, Sept 2004.

[5] Zaki, M. J. (2000). Scalable algorithms for association mining. Knowledge and Data Engineering, IEEE Transactions on, 12(3), 372-390.

[6] Han, J., Pei, J., Yin, Y., & Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data mining and knowledge discovery, 8(1), 53-87.

[7] Yan, X., Zhang, C., & Zhang, S. (2009). Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. Expert Systems with Applications, 36(2), 3066-3076.

[8] Guo, H., & Zhou, Y. (2009, October). An Algorithm for Mining Association Rules Based on Improved Genetic Algorithm and its Application. In Genetic and Evolutionary Computing, 2009. WGEC'09. 3rd International Conference on (pp. 117-120). IEEE.

[9] wang,M., zou,Q. and lin, C. : *Multi dimensions association rules mining on adaptive genetic algorithm*, international conference on uncertainly reasoning on knowledge engineering,IEEE (2011).

[10] Djenouri, Y., Drias, H., Habbas, Z., & Mosteghanemi, H. (2012, December). Bees Swarm Optimization for Web Association Rule Mining. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2012 IEEE/WIC/ACM International Conferences on (Vol. 3, pp. 142-146). IEEE.

[11] Djenouri, Y., Drias, H., & Chemchem, A. (2013, August). A hybrid Bees Swarm Optimization and Tabu Search algorithm for Association rule mining. In Nature and Biologically Inspired Computing (NaBIC), 2013 World Congress on (pp. 120-125). IEEE.

[12] Guvenir H.A., Uysal I. (2000) http://funapp.cs.bilkent.edu.tr. Bilkent University Function Approximation Repository

[13] Goethals B. (2004) http://fimi.ua.ac.be/ Frequent Itemset Mining Implementations Repository

[14] Rupesh Dewang et al. A New Method for Generating All Positive and Negative Association Rules,International Journal on Computer Science and Engineering (IJCSE), ISSN: 0975-3397 Vol. 3 No. 4 Apr 2011.

[15] Anandhavalli M. Optimized association rule mining using genetic algorithm, Advances in Information Mining, ISSN: 09753265, Volume 1, Issue 2, 2009, pp-01-04

[16] Gorawski, M., & Stachurski, K. (2006, April). On Efficiency and Data Privacy Level of Association Rules Mining Algorithms within Parallel Spatial Data Warehouse. In ARES (pp. 936-943).

[17] Gorawski, M., & Siedlecki, Z. (2011). Implementation, optimization and performance tests of privacy preserving mechanisms in homogeneous collaborative association rules mining. In On the Move to Meaningful Internet Systems: OTM 2011 (pp. 347-366). Springer Berlin Heidelberg.