Ranking and selecting association rules based on dominance relationship *

Slim Bouker[1,2,3], Rabie Saidi[1,2], Sadok Ben Yahia[3,4], and Engelbert Mephu Nguifo[1,2]

[1]Clermont Université, Université Blaise Pascal, LIMOS, BP 10448, F-63000 Clermont-Ferrand, France
[2]CNRS, UMR 6158, LIMOS, F-63173 Aubière, France
[3]Faculté des sciences de Tunis, Départment des Sciences de l'Informatique, 1060 Tunis, Tunisie
[4]Départment des Sciences de l'Informatique, Télécom SudParis, UMR CNRS Samovar, 91011 Evry Cedex, France

31 mai 2013

Abstract

The huge number of association rules represents the main hamper that a decision maker faces. In order to bypass this hamper, an efficient selection of rules has to be performed. Since selection is necessarily based on evaluation, many interestingness measures have been proposed. However, the abundance of these measures gave rise to a new problem, namely the heterogeneity of the evaluation results and this created confusion to the decision. In this respect, we propose a novel approach to discover interesting association rules without favoring or excluding any measure by adopting the notion of dominance between association rules. Our approach bypasses the problem of measure heterogeneity and unveils a compromise between their evaluations. Interestingly enough, the proposed approach also avoids another non-trivial problem which is the threshold value specification.

Keywords: Association rules selection, Interestingness measures, Dominance relationship.

1 Introduction

Mining association rules is one of the core tasks in data mining research. Since its first formalization in [AIS93], the research on association rules has become very popular among the data mining researchers. Indeed mining association rules provides an opportunity to extract relevant and valuable relationship between attributes in transaction databases. Currently, association rules are widely used in the *decision making* related to various areas such as telecommunication networks, market and risk management, inventory control etc [Man97]. However, it is well known that data mining algorithms produce an overwhelming number of rules. Hence, the decision maker is unable to determine the most interesting ones and consequently unable to make decisions. In order to face this obstacle, an efficient evaluation of rules has become a compelling need rather than being a rational choice. Several works have been devoted to the study of the interestingness of association rules [HH03], [VLL04]. As a consequence, a panoply of statistical measures, obeying different semantics, have been proposed. Although these measures allow evaluating rules from various sights, yet their abundance (≈ 60) has yielded another problem for the decision maker. Indeed, the outputs of evaluations vary from a measure to another one and may even be contradictory since the measures evaluate differently the rules. That is why, it is common that a given rule be considered relevant according to a measure and irrelevant with respect to another one.

The problem caused by the abundance of measures has led to a trend of works that focuss on proposing approaches to assist the user in selecting the measures qualified to be the most adequate to the decision scope. These approaches can be classified into two main categories namely the expert-based approaches

*This paper is published in the proceeding of the IEEE 24th International Conference on Tools with Artificial Intelligence.

and the property-based ones. In the first category, different studies have compared the ranking of rules by human experts to that yield by various measures. Then, they suggested choosing the measure that yields the closest one to the expert ranking [OSKY04], [TKS02]. These studies were based on specific datasets and experts. Thus, their results cannot be taken as general conclusions. Moreover, in a real problem, it is not always possible to easily get expert's ranking. As for the second category, to reduce the number of measures, many properties have been reported in [GH07]. Geng and Hamilton surveyed the interestingness of measures and summarized nine properties to address that issue. Using properties facilitates a general and practical way to automatically identify interesting measures. This trend has been enriched by different other works [BGGB05], [HZ10], [LMVL08], [BGG$^+$] with an additional number of properties. Nevertheless, these properties are not standards. Hence, they do not guarantee selecting only one best measure. Indeed, a wide range of UCI [1] datasets were also used to study the impact of different properties. The results show no single measure can be elected as an obvious winner [HZ10]. Then, in the case of selecting many measures, the problem related to the variety of outputs, mentioned above, persists. In other words, the user cannot proceed towards a unique selection of rules. Whatever one measure is selected or more, nothing guarantees that they are the "best" ones and some better suited measures may be excluded for the simple reason that the used properties do not take into account the specificity of decision context.

Our contribution lies within this scope. In this paper, we introduce a novel approach that aims at discovering interesting association rules without favoring or excluding any mesure among the used measures. For this purpose, we integrate into the rule selection process, the *skyline operator* [BKS01] whose fundamental principle relies on the notion of *dominance*. The skyline operator is used to resolve mathematical and economics problems such as maximum vectors [KLP75], Pareto set [Mat91] and multi-objective optimization [Ste86]. On the other hand, the skyline operator has received considerable attention in database community and several algorithms, based on block nested loops [BKS01], divide-and-conquer search [KRR02] and index scanning [TEO01], have been developed to meet skyline requests that have different contraints in various computational domains. In our work, the skyline operator comprises the rules that are supposed to be the most interesting ones while taking into account several measures. The dominance relationship, which is the corner stone of the skyline operator, is applied on rules and can be presented as follows: a rule r is said *dominated* by another one r', if for all used measures, r is less relevant than r'. The former rule (*i.e.,* r) is discarded from the result, not because it is not relevant for one of the mesure but because it is not interesting according to the combination of all measures. Our approach bypasses the problem of measure selection by finding a compromise between the different outputs and also bypasses another nontrivial problem which is the threshold value specification. We note that the notion of skyline operator is used in [SRPC11] for mining undominated patterns with respect to a set of measures M. It introduces the notion of Skylineability between measures based on the fact when a pattern p grows (i.e, by adding an item to p), the value of some measures increase or decrease while it remains constant for others. Using this notion of Skylineability, the authors shows how to identify a smaller subset from M which allows for the computation of all undominated patterns with respect to M . This approach can be only applied to a particular kinds of patterns like itemsets, sequential patterns or subgraphs. However, association rules are a combined form (premise and conclusion) of patterns (itemsets) where the evaluation measures are based on the link between premise and conclusion.

The remainder of this paper is organized as follows. Section 2 gives a brief definitions related to association rules and introduces the dominance relationship. We propose and detail our approach of rule selection in section 3. An extension of our approach to enable rule ranking is presented in section 4. Results of the experiments carried out on several datasets are reported in section 5. Concluding points and avenues of future work are sketched in section 6.

2 ASSOCIATION RULES AND DOMINANCE RELATIONSHIP

In this section, we first recall basic definitions related to association rules. Then, we present these rules as numeric vectors within the same dimension after having been evaluated by a set of measures. This vector format, allows us to benefit from the concept of *dominance* and adapt it to our scope as described in section 2.2.

1. http://archive.ics.uci.edu/ml/

	a	b	c	d
t_1			×	×
t_2	×			
t_3	×			×
t_4			×	
t_5		×		×
t_6	×			×
t_7			×	
t_8				×
t_9		×	×	
t_{10}			×	×

(a) *A transaction dataset* $\mathcal{D}$

Rule	*Freq*	*Conf*	*Pearl*
r_1: $a \rightarrow d$	0.20	0.67	0.02
r_2: $b \rightarrow c$	0.10	0.50	0.00
r_3: $b \rightarrow d$	0.10	0.50	0.02
r_4: $c \rightarrow d$	0.20	0.40	0.10
r_5: $d \rightarrow a$	0.20	0.33	0.02
r_6: $d \rightarrow c$	0.20	0.33	0.10
r_7: $c \rightarrow b$	0.10	0.20	0.01
r_8: $d \rightarrow b$	0.10	0.17	0.02

(b) *A table relation* $\Omega(\mathcal{R}, \mathcal{M})$

Name	**Definition**	**Domain**
Frequency	$\frac{supp(X \cup Y)}{\lvert D \rvert}$	**[0, 1]**
Confidence	$\frac{supp(X \cup Y)}{supp(X)}$	**[0, 1]**
Pearl	$\frac{supp(X)}{\lvert D \rvert} \times \lvert \frac{supp(X \cup Y)}{supp(X)} - \frac{supp(Y)}{\lvert D \rvert} \rvert$	**[0, 1]**

(c) *Some measures of* $\mathcal{M}$

Table 1: Example of a dataset transaction and measures.

2.1 Association rules

Let $\mathcal{I}$ be a set of literals called *items*, an itemset corresponds to a non null subset of $\mathcal{I}$. These itemsets are gathered together in the set $\mathcal{L} : \mathcal{L} = 2^{\mathcal{I}} \backslash \emptyset$. In a transactional dataset, each transaction contains an itemset of $\mathcal{L}$. Table 1(a) sketches a transactional dataset $\mathcal{D}$ where 10 transactions, denoted by $t_1, \ldots, t_{10}$ described by 4 items denoted by a, b, c, d. The support of an itemset X, denoted *supp*(X), is the number of transactions containing X.

An association rule r is a relation between itemsets of the form r: X→Y where X and Y are itemsets, and $X \cap Y = \emptyset$. Itemsets X and Y are called, respectively, premise and conclusion of r. The support of r is equal to the number of transactions containing both X and Y, *supp*(r)= *supp*(X∪Y). We notice that interesting measures for association rules are usually defined using support counts as presented in Table 1(b).

2.2 Dominance relationship

After mining association rules from a transactional dataset $\mathcal{D}$ (*e.g.,* Table 1(a)), a set $\mathcal{R}$ of rules is obtained (*e.g.,* Table 1(b) first column). Rules of $\mathcal{R}$ are evaluated with respect to a set $\mathcal{M}$ of measures (*e.g.,* Table 1(c)) to form a relational table Ω (*e.g.,* Table 1(b)). Formally, $\Omega = (\mathcal{R}, \mathcal{M})$ with the set $\mathcal{M} = \{m_1, \ldots, m_k\}$ of measures as attributes and the set $\mathcal{R} = \{r_1, \ldots, r_n\}$ of rules as objects. We denote by $r[m]$ the value of the measure m for the rule r, $r \in \mathcal{R}$ and $m \in \mathcal{M}$. Since the evaluation of rules varies from a measure to another one, using several measures could lead to different outputs (relevant rules with respect to a measure). For example, r_1, and r_2 are the best two rules with respect to the *Confidence* measure whereas it is not the case according to the evaluation of *Pearl* measure which favors r_4 and r_6. This difference of evaluations is confusing for any process of rule selection or ranking.

Based on the above formulation of Ω, we can utilize the notion of dominance between rules to address their ranking as well as the selection of relevant ones. Before, formulating the dominance relationship between rules we need to define it at the level of measure values. To do that, we define value dominance as follows:

Definition 1 *(Value Dominance) Given two values of a measure m corresponding to two rules r and r', we say that $r[m]$ dominates $r'[m]$, denoted by $r[m] \succeq r'[m]$, iff $r[m]$ is preferred to $r'[m]$. If $r[m] \succeq r'[m]$ and $r[m] \neq r'[m]$ then we say that $r[m]$ strictly dominates $r'[m]$, denoted $r[m] \succ r'[m]$.*

To make the dominance relationship scale to the level of rules, we give the following definition:

Definition 2 *(Rule Dominance) Given two rules r, $r' \in \mathcal{R}$, the dominance relationship according to the set of measures $\mathcal{M}$ is defined as follows:*

- *r dominates r', denoted $r \succeq r'$, iff $r[m] \succeq r'[m]$, $\forall\, m \in \mathcal{M}$.*
- *If $r \succeq r'$ and $r' \succeq r$,* i.e., *$r[m] = r'[m]$, $\forall\, m \in \mathcal{M}$ then r and r' are said **equivalent**, denoted $r \equiv r'$.*
- *If $r \succeq r'$ and $\exists\, m \in \mathcal{M}$ such that $r'[m] \succ r[m]$, then r' is **strictly dominated** by r and we note $r \succ r'$.*

It is easy to check that the strict dominance relationship fulfils the following properties:

- **irreflexive**: $r \not\succ r$, *i.e,* $r \succ r$ is false for each $m \in \mathcal{M}$,
- **transitive**: $\forall\, r$, r' and $r'' \in \mathcal{R}$, if $r \succeq r'$ and $r' \succeq r''$ then $r \succeq r''$.

Example 1 *Given the relation table Ω in Table 1(b), the rule r_3 strictly dominates r_2 since $r_3[Freq] \succeq r_2[Freq]$, $r_3[Conf] \succeq r_2[Conf]$ and $r_3[Pearl] \succ r_2[Pearl]$.*

Whenever a rule r dominates another one r' with respect to $\mathcal{M}$, this means that r is equivalent to or

better than r' for all measures. Hence, the dominance relationship allows comparing concurrently two rules with respect to all measures. Hence, it can be used to bypass the problem of difference of evaluations. Rules dominated by other ones (at least one), according to $\mathcal{M}$, are not relevant and have to be eliminated. The skyline operator for association rules formalizes this intuition.

Definition 3 *(Skyline operator) The skyline of Ω over $\mathcal{M}$, denoted by $Sky_M(\Omega)$, is the set of rules from Ω defined as follows:*

$$Sky_{\mathcal{M}}(\Omega) = \{\ r \in \mathcal{R} \mid \nexists\ r' \in \mathcal{R},\ r' \succ r\}$$

In other words, the skyline of Ω is the set of undominated rules of $\mathcal{R}$ with respect to $\mathcal{M}$. For instance, from the relation table Ω in Table 1(b), $Sky_M(\Omega) = \{r_1, r_4\}$ since there is no rule in $\mathcal{R}$ dominating r_1 or r_4.

3 DISCOVERING UNDOMINATED RULES

To discover undominated rules, we adopt the principle of approaches oriented divide-and-conquer search [KRR02] used for answering queries in database applications. In the following, we introduce the necessary formalization that would be of need for the generation of the undominated rules. Based on this formalization we propose an algorithm, called SkyRule, that puts the skyline operator.

3.1 Formalization

To discover the undominated rules, a naïve approach consist in comparing each rule with all other ones. However, association rules are often present in huge number which make it very costly to perform pairwise comparisons. In the following, we show how to remedy this problem. First, we introduce the notion of *reference rule*.

Definition 4 *(Reference Rule) A reference rule $r^{\perp}$ is a fictitious rule that dominates all the rules of $\mathcal{R}$. Formally: $\forall\ r \in \mathcal{R}$, $r^{\perp} \succeq r$.*

Example 2 *From the relational table Ω given in Table 1, we can consider $r^{\perp}$ as the fictitious rule such that for each measure $m \in \mathcal{M}$, $r^{\perp}[m]$ is the maximal value appearing in the active domain of m, i.e., $r^{\perp} = \langle 0.2, 0.67, 0.10\rangle$. Hence, it does not exist any rule in $\mathcal{R}$ that dominates $r^{\perp}$.*

In practice, measures are heterogenous and defined within different domains. For our purpose, $\mathcal{M}$ have to be normalized into $\widehat{\mathcal{M}}$ within one interval $[p,q]$. In other words, each measure $m \in \mathcal{M}$ must be normalized into $\widehat{m} \in \widehat{\mathcal{M}}$ within $[p,q]$. The normalization of a given measure m is performed depending on its domain and the statistical distribution of its active domain. We recall that the active domain of a measure m is the set of its values in Ω. The normalization is a statistical problem which is beyond the scope of this paper. Worth of mention, the normalization of a measure does not modify the domination relationship between two given values.

Definition 5 *(Degree of similarity) Given two rules r, $r' \in \mathcal{R}$, the degree of similarity between r and r' with respect to $\widehat{\mathcal{M}}$ is defined as follows:*

$$DegSim(r, r') = \frac{\sum_{i=1}^{k} \mid r[\widehat{m}_i] - r'[\widehat{m}_i] \mid}{k}$$

with $\mid x - y \mid$ is the absolute value of $(x - y)$, x and $y \in [p,q]$ and $k = \mid \widehat{\mathcal{M}} \mid$.

Example 3 *Let's consider our running example using the relation table Ω in Table 1(b). Since all measures are defined within the same domain [0,1], we can compute, without normalization, the degree of similarity between each rule and the reference rule given in the previous example. $DegSim(r^{\perp},r_1)$ = 0.02, $DegSim(r^{\perp},r_2)$ = 0.12, $DegSim(r^{\perp},r_3)$ = 0.11, $DegSim(r^{\perp},r_4)$ = 0.09, $DegSim(r^{\perp},r_5)$ = 0.14, $DegSim(r^{\perp},r_6)$ = 0.11, $DegSim(r^{\perp},r_7)$ = 0.22, $DegSim(r^{\perp},r_8)$ = 0.23.*

After giving the necessary definitions (reference rule and degree of similarity), the following lemma gives a remedy to the issue evoked in the beginning of section 3.1. Indeed, it offers a swifter solution rather than pairwise comparisons; to find undominated rules.

Lemma 1 *Let $r \in \mathcal{R}$ be a rule having the minimal degree of similarity with respect to $r^{\perp}$, then $r \in Sky_{\mathcal{M}}(\Omega)$.*

Proof 1 *Let $r \in \mathcal{R}$ be a rule having the minimal degree of similarity with respect to $r^{\perp}$ and we suppose that $r \notin Sky_M(\Omega)$, then there exists a rule $r' \in \mathcal{R}$ that strictly dominates r, which means that $\forall\ m \in \mathcal{M}$, $r'[m] \succeq r[m]$ and $\exists\ m' \in \mathcal{M}$, $r'[m'] \succ r[m']$. Hence, we have $DegSim(r^{\perp},r') < DegSim(r^{\perp},r)$. The latter inequivalent contradicts our hypothesis, since r has the minimal degree of similarity with respect to $r^{\perp}$.*

After identifying an undominated rule r, the rules dominated by r must be identified by comparing them to r. Naïvely, r must be compared to all rules in $\mathcal{R}$, yet we show in the following that we can even reduce the set of rules to be compared with r into a subset of $\mathcal{R}$.

Definition 6 *(undominated space) Let r be an undominated rule. If there exists a rule r' which is not dominated by r such that $r \not\equiv r'$, then there exists at least a measure $m \in \mathcal{M}$ such that $r'[m] \succ r[m]$. Since there exist k measures in $\mathcal{M}$, then there are k sets such that each one of them may contain rules not dominated by r. For each measure $m_i \in \mathcal{M}$, i=1,...,k, the corresponding set s_i^r of rules which are not dominated by r is defined as follows:*

$$s_i^r = \{\ r' \in \mathcal{R} \mid r \not\succ r' \text{ and } r'[m_i] \succ r\ [m_i]\}$$

These k sets compose the undominated space of r, denoted $\mathcal{S}^r = \{s_i^r\}$, i=1,...,k.

Example 4 *From our toy example presented in Table 1(b), for the undominated rule r_1, $s_1^{r_1} = \emptyset$, we have $s_2^{r_1} = \emptyset$ and $s_3^{r_1} = \{r_4, r_6\}$. $s_1^{r_1}$ and $s_2^{r_1}$ are empty since there is no rule $r \in \mathcal{R}$ such that $r[m_1] \succ r_1[m_1]$ or $r[m_2] \succ r_1[m_2]$. However, $s_3^{r_1}$ contains r_4 and r_6 since $r_4[m_3] \succ r_1[m_3]$ and $r_6[m_3] \succ r_1[m_3]$. Following a similar reasoning, for the undominated rule r_4, we have $s_1^{r_4} = \emptyset$, $s_2^{r_4} = \{r_1, r_2, r_3\}$ and $s_3^{r_4} = \emptyset$.*

Lemma 2 *Let $r, r' \in \mathcal{R}$ be two undominated rules and $s^r \in \mathcal{S}^r$. If $r' \notin s^r$, then $\forall\ r'' \in s^r$, $r' \not\succ r''$.*

Proof 2 *Given $r, r' \in \mathcal{R}$ two undominated rules and $s^r \in \mathcal{S}^r$ corresponding to a measure $m \in \mathcal{M}$. If $r' \notin s^r$, then $r'[m] \not\succ r[m]$ which means that $r[m] \succeq r'[m]$ (1). Moreover, since $r'' \in s^r$ then $r''[m] \succ r[m]$ (2). According to the dominance transitivity, (1) and (2) lead to $r''[m] \succ r'[m]$. Hence, $r' \not\succ r''$.*

Lemma 3 *Let be $r, r' \in \mathcal{R}$ and $s^r \in \mathcal{S}^r$ such that r is an undominated rule and $r' \in s^r$. If r' has the minimal degree of similarity with respect to $r^\perp$ among the rules in s^r, then $r' \in Sky_M(\Omega)$.*

Proof 3 *Given $r, r' \in \mathcal{R}$ and $s^r \in \mathcal{S}^r$ such that $r' \in s^r$ and r' has the minimal degree of similarity with $r^\perp$ among the rules in s^r. Suppose that $r' \notin Sky_M(\Omega)$, then it means that there exists a rule $r'' \in \mathcal{R}$ such that $r'' \succ r'$. According to lemma 2, r'' must be in s^r since any rule not belonging to s^r cannot dominate r'. Moreover, $\forall\ m \in \mathcal{M}$, $r''[m] \succeq r'[m]$ and $\exists\ m' \in \mathcal{M}$, $r''[m'] \succ r'[m']$. Hence, $DegSim(r^\perp, r'') < DegSim(r^\perp, r')$ which contradicts our hypothesis since r' has the minimal degree of similarity with $r^\perp$ in s^r.*

3.2 SkyRule Algorithm

Based on the formalization, we proposed the SkyRule algorithm allowing to discover undominated rules. In SkyRule algorithm, we use the following variables for accumulating data during the execution of the algorithm:

- The variable *Sky*: is a variable initialized to empty set, it is used to keep track of the undominated rules.
- The variable C: is a variable that contains the set of all current candidate rules to be qualified as undominated; it is initialized to $\mathcal{R}$.
- The variable $\mathcal{E}$: is a variable that contains all current set covering the undominated space of all undominated rules; it is initialized to $\mathcal{R}$ since initially, all rules are considered undominated.

Algorithm 1: SkyRule

```
Input: Ω = (R, M)
Output: Sky: set of undominated rules of Ω.
1  Begin
2    Sky ← ∅
3    C ← R
4    E ← {R}
5    While C ≠ ∅ do
6      r* ← r ∈ C having min(DegSim(r,r⊥))
7      C ← C\{r*}
8      for i=1 to k do
9        s_i^r* ← ∅
10     Sky ← Sky ∪ {r*}
11     Foreach e ∈ E such that r* ∈ s do
12       Foreach r ∈ e do
13         If r* ≻ r then
14           C ← C\{r}
15         Else
16           for i=1 to k do
17             If r[m_i] ≻ r*[m_i] then
18               s_i^r* ← s_i^r* ∪{r}
19       E ← E\{e}
20     E ← E ∪ {s_1^r*, . . . , s_k^r*}
21   return Sky
22 End
```

Informally, the algorithm works as follows:

- If the set of candidate rules C is empty, then the algorithm terminates and all undominated rules are outputted through the variable *Sky*.

- Otherwise, each rule r in C might be an undominated one. If r has the minimal degree of similarity with the reference rule $r^{\perp}$, then r is an undominated rule and is added to Sky (*i.e.,* r is no longer candidate and is deleted from C). After that, only the undominated space containing r is explored as follows: for each rule r', in this undominated space, is compared with r. Two cases have to be distinguished:
  1. if r' is dominated by r, then r is no longer candidate and it is withdrawn from C.
  2. otherwise, r' is not dominated by r, *i.e.,* r' is still a candidate rule and it is added to the undominated subspace of r according to definition 6.

  Then, the undominated space containing r is deleted from $\mathcal{E}$ and the undominated space of r is added to $\mathcal{E}$. This process comes to an end when all candidates are handled.

4 RANKING ASSOCIATION RULES

The SkyRule algorithm allowed identifying the undominated rules which are supposed to be the most relevant ones. However, this output might not be enough answer to a personalized user query. Indeed, the user often needs a specified number of relevant rules which may be more or less than what the SkyRule algorithm generates. In the first case *i.e.,* the user asks for a subset of the undominated rules, a selection is required among the SkyRule output. Since, SkyRule generates only relevant rules, the most relevant among them must be returned to the user. This selection cannot be performed unless a ranking has been done within the undominated rules. In the second case *i.e.,* the user asks for a set of relevant rules larger than the set of undominated rules, the rules that must be added to the SkyRule output are necessarily a part from the set of dominated rules. The composition of this part requires a selection among all the dominated rules. This selection cannot be performed unless a ranking has be carried within the dominated rules. Hence, a ranking process must be performed on the whole set of rules.

In the remainder, we introduce our second contribution: we show that we can perform a comprehensive ranking using SkyRule. For this purpose, we give the two following objective conditions:

1. Any dominated rule cannot be better ranked than an undominated one.
2. Two undominated rules must be ranked based on degree of similarity with respect to reference rule.

4.1 Succession relationship

In the following, we introduce the notion of *succession relationship*. This notion is based on the dominance relationship. First, we define it at the level of rules. Then, we define it at the level of rule sets. Both definitions are essential to state Lemma 4. That lemma puts the corner stone of our approach that uses the skyline operator to establish a ranking process. This process is described by RankRule (*c.f.,* Algorithm 2).

Definition 7 *(Successor rule) Let's consider two rules $r, r' \in \mathcal{R}$, we say that r succeeds r', denoted by $r \lhd r'$ iff $r' \succ r$ and $\nexists\, r''$ such that $r' \succ r'' \succ r$.*

Example 5 *Consider the relation table Ω in Table 1(b), then we have $r_6 \lhd r_4$ but $r_5 \not\lhd r_4$ since $r_4 \succ r_6 \succ r_5$.*

Definition 8 *(Succession Operator) Let E be a set of rules such that $E \subseteq \mathcal{R}$. The successeur set of E in $\mathcal{R}$ with respect to $\mathcal{M}$ is defined as follows: $Succ_{\mathcal{M}}(E,\mathcal{R})$ $= \{\, r \in \mathcal{R} \setminus E \mid \exists\, r' \in E,\, r \lhd r' \wedge \nexists\, r'' \in E,\, (r'' \succ r \wedge r \not\lhd r'')\}$*

Example 6 *Let's consider our running example using the relation table Ω in Table 1(b) and suppose $E = \{r_1, r_4\}$. Then, we have $r_1 \succ r_3 \succ r_2$, $r_1 \succ r_5 \succ r_7$, $r_5 \succ r_8$ and $r_4 \succ r_6 \succ r_5$ then $Succ_{\mathcal{M}}(E,\mathcal{R}) = \{r_3, r_6\}$. Notice that, although $r_5 \lhd r_1$, $r_5 \notin Succ_{\mathcal{M}}(E,\mathcal{R})$ since $r_5 \not\lhd r_4$.*

Lemma 4 *Given a set of rules $E \subseteq \mathcal{R}$, the following relation is fulfilled:*

$$Succ_{\mathcal{M}}(Sky_{\mathcal{M}}(E),E) = Sky_{\mathcal{M}}(E \backslash Sky_{\mathcal{M}}(E\ ))$$

Proof 4 *Let E be a set of rules, such that $E \subseteq \mathcal{R}$:*
1. First we have to show that $Succ_{\mathcal{M}}(Sky_M(E),E) \subseteq Sky_M(E \backslash Sky_M(E))$:
Given, $r \in Succ_{\mathcal{M}}(Sky_{\mathcal{M}}(E),E)$ then $r \in E \backslash Sky_{\mathcal{M}}(E)$. For all $r' \in Sky_{\mathcal{M}}(E)$, two cases have to be distinguished:

- *If $r' \succ r$, then $r \lhd r'$ which means that $\nexists\, r'' \in E \backslash Sky_{\mathcal{M}}(E)$ such that $r' \succ r'' \succ r$.*
- *If $r' \not\succ r$, then $\nexists\, r''$ in $E \backslash Sky_{\mathcal{M}}(E)$ such that $r' \succ r''$ and $r'' \succ r$*

Thus, r cannot be dominated by any rule in $E \backslash Sky_{\mathcal{M}}(E)$ i.e., $r \in Sky_{\mathcal{M}}(E \backslash Sky_{\mathcal{M}}(E))$.
2. Second, we have to show that $Succ_{\mathcal{M}}(Sky_{\mathcal{M}}(E),E) \supseteq Sky_{\mathcal{M}}(E \backslash Sky_{\mathcal{M}}(E\ ))$:
Given $r \in Sky_{\mathcal{M}}(E \backslash Sky_{\mathcal{M}}(E\ ))$ then $\nexists\ r' \in$

$E\backslash Sky_{\mathcal{M}}(E)$ such that $r' \succ r$ ***(a)****. Moreover, since $r \in E\backslash Sky_{\mathcal{M}}(E)$ then $\exists\, r'' \in Sky_{\mathcal{M}}(E)$ such that $r'' \succ r$* ***(b)****. Thus,* ***(a)*** *and* ***(b)*** *leads to that $r \lhd r''$* ***(c)****. Furthermore, we suppose that $\exists\, r' \in Sky_{\mathcal{M}}(E)$ such that $r_1 \succ r$ and $r \not\lhd r_1$, then $\exists\, r_2 \in E\backslash Sky_{\mathcal{M}}(E)$ such that $r_1 \succ r_2 \succ r$ which contradicts our hypothesis (see* ***(a)****). Thus, $\nexists\, r_2 \in E\backslash Sky_{\mathcal{M}}(E)$ such that $r_1 \succ r_2 \succ r$* ***(d)****. Hence, according to* ***(c)*** *and* ***(d)****, r belongs to $Succ_{\mathcal{M}}(Sky_{\mathcal{M}}(E),E)$.*

Algorithm 2: RankRule

Input: $\Omega = (\mathcal{R}, \mathcal{M})$
Output: Ordered sets of ordered rules

```
Begin
    p ← 0
    While R ≠ ∅ do
        p ← p + 1
        E_p ← SkyRule(Ω)
        R ← R\E_p
        Ω ← (R, M)
    return (E_1, . . ., E_p)
End
```

Example 7 *In this example, we apply the* RankRule *algorithm on Ω of Table 1(b). Since, both r_1 and r_4 are undominated rules then $E_1 = \{r_1, r_4\}$. Now, we ignore r_1 and r_4, the rules which are not dominated are r_3 and r_6. In fact, r_3 is only dominated by r_1 and r_6 is only dominated by r_1, then $E_2 = \{r_3, r_6\}$. Now we also ignore r_3 and r_6, the rules which are not dominated are r_2 and r_5. In fact, r_2 is dominated by r_3 and r_5 is only dominated by r_6, then $E_3 = \{r_2, r_5\}$. Finally, we have $E_4 = \{r_7, r_8\}$. This example is illustrated by Figure 1. The arrow indicates the process direction starting from the undominated rules. E_1 contains the top ranked rules which are themselves ranked within E_1 from left to right based on DegSim: r_1 is better ranked than r_4.*

4.2 Duality

The RankRule algorithm performs ranking by starting from the set of the most relevant rules (*i.e.,* the undominated rules). The latter are then used to identify the next ranked set (*i.e.,* the successor). Nevertheless, another dual possibly remains explorable. It relies on starting from the set of the less relevant rules (*i.e.,* rules that do not dominate other ones) and use them to identify the previous ranked rule set that we called it *predecessor* set. A complete formalization of this dual perspective is beyond the scope of this paper. Nevertheless, we explain how it works by the following illustrative example.

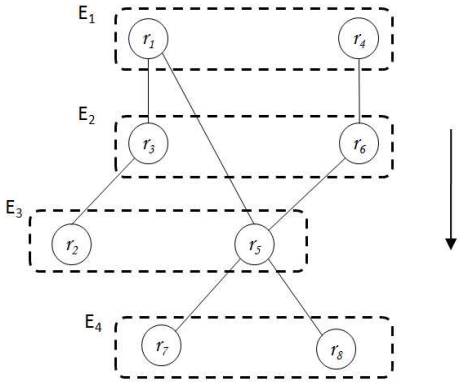

Figure 1: The output of RankRule applied on Ω given by table 1(b).

Example 8 *We consider Ω of Table 1 (b). First, we identify the set of rules which do not dominate any other rules. These rules are r_2, r_7 and r_8 then we have $E_4 = \{r_2, r_7, r_8\}$. Now we have to ignore these rules. The rules which do not dominate any other rules are r_3 and r_5. In fact, r_3 dominates only r_2 and r_5 dominates only r_7 and r_8, then $E_1 = \{r_3, r_5\}$. Now we also ignore r_3 and r_5, The rules that do not dominate any other rules are r_1 and r_6 since they dominate r_3 and r_5 respectively, then $E_2 = \{r_1, r_6\}$. Finally, $E_1 = \{r_4\}$.*

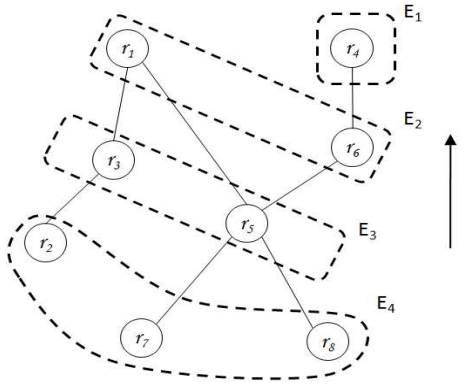

Figure 2: The dual RankRule applied on Ω given table 1(b).

5 EXPERIMENTAL STUDY

In this section, our objectives are at number of two. First, we show through extensive experiments that SkyRule provides interesting instance reduction compared to the initial set of rules. Second, we assess whether the number of measures has any uniform impact on the number of undominated rules. These experiments were carried out on benchmark datasets taken from the UCI Machine Learning Repository. Table 2 summarizes the characteristics of these datasets. All the tests were performed on a 1.73 GHz Intel processor with Linux operating system and 2 GB of RAM memory.

Dataset	♯ *items*	♯ *transactions*	*Avg. size of transactions*
Diabete	75	3196	37
Flare	39	1389	10
Iris	119	8124	23
*Monks*1	19	124	7
*Monks*2	19	169	7
*Monks*3	19	122	7
Nursery	32	12960	9
Zoo	42	101	9

Table 2: Benchmark dataset characteristics.

5.1 Reduction of number of rules

In this subsection, we show the ability of our approach to considerably reduce the huge numbers of rules generated from our experimental datasets. Our experiments batch aims to compare our approach to another one based on thresholds. For this purpose, we assign for each measure $m \in \mathcal{M}$, a threshold ε_m such that ε_m is the minimum value of the skyrules with respect to m, *i.e.,* $\varepsilon_m = min\{r[m] \mid r \in Sky_M(\Omega)\}$. This ensures that all undominated rules will be generated from the algorithm based on thresholds. For instance, in our running example (*c.f.,* Table 1(b)), $\varepsilon_{freq} = 0.10$, $\varepsilon_{conf} = 0.17$ and $\varepsilon_{pearl} = 0.00$. The set of resulting rules is called the threshold-based rules denoted by *TB rules*. These experiments have the benefit of quantifying the reduction of rules brought by SkyRules in the case where a user is able to perfectly specify thresholds for mining association rules algorithm based on thresholds. Hence, we compare the number of undominated rules with respect to that of *TB rules* and the total number of association rules (denoted *A-R*). We considered a number of combinations of measures: Confidence [AIS93], Recall [LFZ99], Pearl [Pea88], Loevinger [Loe47], Zhang [Zha00].

For each set of measures, Table 3 compares the size of undominated rules versus that of *TB rules* and that of all association rules. The goal is to illustrate the problem of the huge number of association rules even with threshold-based algorithm which makes difficult to discover interesting ones. In contrast, the number of undominated rules is always low and does not exceed 9784. Interestingly, the gain of a undominated rules is always important (very high in almost all datasets). Table 4 summarizes this result by sketching, for each set of measures, the minimal/average/maximal number of undominated rules, the average number of *TB rules* and the average gain of undominated rules versus the *TB rules*. The average gain rate is measured as follows: $\frac{size\ of\ TB\ rules}{size\ of\ Sky\text{-}R}$.

5.2 Impact of measure variation on the number of rules

In what follows, we put the focus on the evolution of the undominated rules cardinalities with respect to measure variation. Table 3 shows the effect of variation of $\mathcal{M}$ on undominated rules, *TB rules* and all rules. We can notice that the number of all rules is obviously constant. In contrast, the number of *TB rules* is sensitive to the variation of cardinality of $\mathcal{M}$. Indeed, by adding each time a measure to $\mathcal{M}$, the number of *TB rules* decreases. However, the number of undominated rules may decrease or increase. The decrease can be explained by the fact that an association rule can be undominated with respect to a set of measure M_1 and dominated with respect to M_2, such that $M_1 \subset M_2$. For example, if two rules r and r' are equivalent and undominated with respect to M_1, there is a possibility that one of them dominates the other by considering one more measure. On the other hand, the increase can be explained by the fact that an association rule can be dominated with respect to M_1 and undominated with respect to M_2. For example, consider a rule r which dominates another r' with respect to M_1, by adding a mesure m to M_1, such that $r'[m] \succ r[m]$, then r' is no longer dominated by r.

6 CONCLUSION

In this paper, we introduced an approach that addresses the problem of rule selection and ranking. This approach is not hindered by the abundance of measures which is the issue of several works. These works have been devoted to measure selection in order to find one best measure, whereas the real issue lies in select-

Table 3: Undominated rules *vs* TB rules and All rules

Datasets (min_{freq} %)		{Conf; Loev}	{Conf; Pearl}	{Conf; Recall}	{Conf; Zhang}	{Conf;Pearl; Recall}	{Conf;Loev; Zhang}	{Conf;Loev;Pearl; Recall;Zhang}
Diabetes (10,00)	Sky-R	3411	9	6651	2996	9	171	171
	TB-R	59314	58124	59206	59309	44813	44602	42126
	A-R	62132	62132	62132	62132	62132	62132	62132
Flare (10,00)	Sky-R	4975	48	4978	4857	48	48	48
	TB-R	56163	57101	56451	54524	53197	53116	52819
	A-R	57476	57476	57476	57476	57476	57476	57476
Iris (0,00)	Sky-R	246	246	246	246	246	246	246
	TB-R	440	440	440	440	440	440	440
	A-R	440	440	440	440	440	440	440
Monks1 (1,00)	Sky-R	768	1	788	656	1	1	1
	TB-R	60417	60692	59418	59452	58904	58811	58327
	A-R	62184	62184	62184	62184	62184	62184	62184
Monks2 (1,00)	Sky-R	279	3	215	202	3	3	3
	TB-R	59611	59702	59568	59544	59103	58917	58662
	A-R	59976	59976	59976	59976	59976	59976	59976
Monks3 (1,00)	Sky-R	1028	2	713	781	4	2	2
	TB-R	58662	58369	57922	58436	57816	57734	56038
	A-R	59304	59304	59304	59304	59304	59304	59304
Nursery (2,00)	Sky-R	497	2	304	342	8	2	2
	TB-R	23872	23901	23875	23417	23176	22806	22139
	A-R	25062	25062	25062	25062	25062	25062	25062
Zoo (10,00)	Sky-R	9784	36	9415	9112	36	36	36
	TB-R	67991	67305	67872	66146	65328	65116	63926
	A-R	71302	71302	71302	71302	71302	71302	71302

Table 4: Gain of the undominated rules

Measures	Average number of Sky-R	Average number of TB-R	Average gain of Sky-R
{Conf;Loev}	2623,50	48308,75	18,41
{Conf;Pearl}	43,37	40908,12	943,23
{Conf;Recall}	2913,75	48094,00	16,50
{Conf;Zhang}	2399,00	47658,50	19,86
{Conf;Loev;Recall}	43,37	45347,12	1045,58
{Conf;Pearl;Zhang}	63,62	45192,75	710,35
{Conf;Loev;Pearl;Recall;Zhang}	63,62	44309,62	696,47

ing and ranking rules to help with decision making. We proposed two algorithms SkyRule and RankRule to perform these two tasks based on the dominance relationship. When using our algorithms, the user does not have to worry neither about the heterogeneity of measures nor about specifying thresholds. On the other hand, experimental results carried out on benchmark datasets showed important profits in terms of compactness of the undominated rules.

An important direction for future work is to take into account similarity between rules in order to eliminate redundant undominated rules. Indeed, similar rules often have the same quality *i.e.,* they have almost identical values for the various measures. Another important task consists on setting up an approach aiming at discovering undominated rules during the phase of the extraction rules which will improve the performance of the SkyRule algorithm. Finally, an impor-

tante prospective is to plan to formalize the dual of RankRule and to find the relationship between them that allows to obtain the output of one of them from the output of the other.

References

[AIS93] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *Proceedings of the ACM SIGMOD, Washington, USA*, pages 207–216, 1993.

[BGG+] R. Belohlávek, D. Grissa, S. Guillaume, E. M. Nguifo, and J. Outrata. Measuring the interestingness of discovered knowledge: A principled approach. *AMAI 2013, to appear*.

[BGGB05] Julien Blanchard, Fabrice Guillet, Régis Gras, and Henri Briand. Using information-theoretic measures to assess association rule interestingness. In *ICDM*, pages 66–73, 2005.

[BKS01] S. Borzsony, D. Kossmann, and K. Stocker. The skyline operator. In *Proceedings of the 17th ICDE*, pages 421–430, Heidelberg, Germany, 2001.

[GH07] Liqiang Geng and Howard J. Hamilton. Choosing the right lens: Finding what is interesting in data mining. In *Quality Measures in Data Mining*, pages 3–24. 2007.

[HH03] Robert J. Hilderman and Howard J. Hamilton. Measuring the interestingness of discovered knowledge: A principled approach. *Intell. Data Anal.*, 7(4):347–382, 2003.

[HZ10] Mojdeh Jalali Heravi and Osmar R. Zaïane. A study on interestingness measures for associative classifiers. In *SAC*, pages 1039–1046, 2010.

[KLP75] H. T. Kung, Fabrizio Luccio, and Franco P. Preparata. On finding the maxima of a set of vectors. *J. ACM*, 22(4):469–476, 1975.

[KRR02] Donald Kossmann, Frank Ramsak, and Steffen Rost. Shooting stars in the sky: An online algorithm for skyline queries. In *VLDB*, pages 275–286, 2002.

[LFZ99] Nada Lavrac, Peter A. Flach, and Blaz Zupan. Rule evaluation measures: A unifying view. In *ILP*, pages 174–185, 1999.

[LMVL08] Philippe Lenca, Patrick Meyer, Benoît Vaillant, and Stéphane Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2):610–626, 2008.

[Loe47] J. Loevinger. *A systemic approach to the construction and evaluation of tests of ability Application.* Psychological monographs, 1947.

[Man97] Heikki Mannila. Methods and problems in data mining. In *ICDT*, pages 41–55, 1997.

[Mat91] Jirí Matousek. Computing dominances in E^n. *Inf. Process. Lett.*, 38(5):277–278, 1991.

[OSKY04] M. Ohsaki, Y. Sato, S. Kitaguchi, and H. Yokoi. Comparison between objective interestingness measures and real human interest in medical data mining. In *Proceedings of the 17th IEA/AIE 2004, Springer-Verlag*, pages 1072–1081, 2004.

[Pea88] Judea Pearl. On logic and probability. *Computational Intelligence*, 4:99–103, 1988.

[SRPC11] Arnaud Soulet, Chedy Raïssi, Marc Plantevit, and Bruno Crémilleux. Mining dominant patterns in the sky. In *ICDM*, pages 655–664, 2011.

[Ste86] R. Steuer. *Multiple Criteria Optimization: Theory, Computation and Application.* (John Wiley, 546), 1986.

[TEO01] Kian-Lee Tan, Pin-Kwang Eng, and Beng Chin Ooi. Efficient progressive skyline computation. In *VLDB*, pages 301–310, 2001.

[TKS02] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th ACM SIGKDD*, pages 32–41, 2002.

[VLL04] Benoît Vaillant, Philippe Lenca, and Stéphane Lallich. A clustering of interestingness measures. In *Discovery Science*, pages 290–297, 2004.

[Zha00] Tao Zhang. Association rules. In *PAKDD*, pages 245–256, 2000.