Intelligent Data Analysis 17 (2013) 195–215
DOI 10.3233/IDA-130574
IOS Press



# Interestingness measures for association rules within groups

Aída Jiménez*, Fernando Berzal and Juan-Carlos Cubero
*Department of Computer Science and Artificial Intelligence, CITIC, University of Granada, Granada, Spain*

**Abstract.** The work described in this paper addresses the study of association rules within groups of individuals. The analysis of the characteristics and the behavior of the individuals belonging to such groups in a given database is powerful in practice, since it provides a mechanism to deal with groups rather than isolated individuals. In this paper, we define group association rules and we study interestingness measures for them. These interestingness measures can be used to rank, not only groups of individuals, but also rules within each group. We also compare the rankings provided by those different interestingness measures in order to determine which one provides a better alternative depending on the kind of situations we wish to highlight within large databases with many different (and overlapping) groups of individuals.

Keywords: Group association rules, interestingness measures



## 1. Introduction

The approach we propose in this paper intends to solve the second-order data mining problem that often arises in practice; i.e. when the results of a data mining process have to be mined themselves due to their huge volume. Researchers in the data mining field have traditionally focused their efforts to obtain fast and scalable algorithms in order to deal with huge amounts of data. When dealing with association rules, for instance, the overwhelming number of discovered rules, usually in the order of thousands or even millions, makes them of limited use in practice. The mere volume of these sets of rules causes the aforementioned second-order data mining problem [3].

Databases can naturally contain groups of individuals that share some characteristics [18]. For example, within a census database, we can find many different groups of individuals, e.g. according to their sex, their marital status, whether they have children, or just by combining several of such attributes. Our proposal consists in automatically identifying potentially useful groups of related association rules and ranking the resulting group association rules so that expert users can more easily sift through vast amounts of association rules.

Our paper is organized as follows. In Section 2, we present the state of the art and describe some rule interestingness measures as proposed in the literature. In Section 3 we adapt some of the interestingness measures that have been defined for association rules in order to group association rules. In Section 4, we explain how to rank groups and group association rules within each group. Section 5 introduces criteria to compare alternative rankings. We analyze the experimental results we have obtained in Section 6. Section 7 provides some guidelines to identify the most interesting groups according to the experiments. Finally, we end our paper with some conclusions in Section 8.

*Corresponding author: Aída Jiménez, Department of Computer Science and Artificial Intelligence, CITIC, University of Granada, Granada 18071, Spain. E-mail: aidajm@gmail.com.

## 2. State of the art

Association rules have been used to analyze co-occurrence relationships among frequent itemsets in transactional and relational databases. Let $I = \{I_1, I_2, \ldots, I_m\}$ be a set of items. Let $D$ be a set of database transactions where each transaction $T$ is a set of items such that $T \subseteq I$. Let $S$ be a set of items. A transaction $T$ is said to contain $S$ if and only if $S \subseteq T$ [10]. An *association rule* is an implication of the form $A \Rightarrow C$, where $A \subseteq I$, $C \subseteq I$, and $A \cap C = \emptyset$.

In the following sections we will present related work to association rules and its interestingness measures as well as a summary of the interestingness measures that have been proposed in order to evaluate association rules.

### 2.1. Related work

One of the challenges when mining patterns and association rules is to deal with the huge amount of them that can be discovered in the mining process. Therefore, it is necessary to establish some kind of measure in order to determine which are the most interesting ones.

This problem has been addressed from several points of views. Several authors resort to statistical techniques in order to evaluate the patterns applying statistical hypothesis tests [22], using empirical Bayes models [6] or log-linear models [23]. Other proposals use clustering techniques to reduce the number association rules. Liu et al. [14] propose the partition of the dataset in clusters of similar transactions and then mine association rules on the summaries of clusters instead of the entire data set. Let et al. [18] instead cluster the resulting association rules according to the values of the attributes involved in the rules. Our proposal is not to cluster the set of association rules but to obtain all the possible groups of transactions in the database. These groups do not have to be disjoint as in a classical clustering, and our goal is to identify the most interesting ones.

In the following section we analyze some of these interestingness measures proposed for association rules that we will adapt and extend to the case of the group association rules.

### 2.2. Interestingness measures for standard association rules

Interestingness measures have been widely used in the data mining area. Tan et al. [20] proposed objective measures to evaluate association patterns. Other measures have been proposed to evaluate association rules [2,4,9,21]. The classical measures used to characterize an association rule are its support and its confidence [1,10].

**Definition 1.** *The support of an itemset $X$ in the database $D$ is defined as the percentage of transactions that contain $X$, i.e.,*

$$supp(X) = P(X).$$

**Definition 2.** *The rule $A \Rightarrow C$ holds in the transaction set $D$ with support $supp(A \Rightarrow C)$, where $supp(A \Rightarrow C)$ is the percentage of transactions in $D$ that contain $A \cup C$, i.e.,*

$$supp(A \Rightarrow C) = supp(A \cup C) = P(A \cup C).$$

**Definition 3.** *The rule $A \Rightarrow C$ has confidence conf $(A \Rightarrow C)$ in the transaction set $D$, where conf $(A \Rightarrow C)$ is the percentage of transactions in $D$ containing $A$ that also contain $C$, i.e.,*

$$conf\,(A \Rightarrow C) = P(C|A) = \frac{supp\,(A \Rightarrow C)}{supp\,(A)}$$

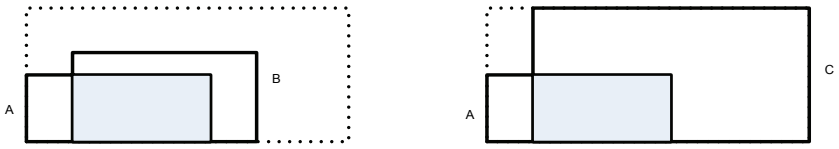



Fig. 1. Graphical depiction of two rules, $A \Rightarrow B$ and $A \Rightarrow C$, both with the same confidence but with different consequent supports. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDA-130574)

Confidence has some drawbacks, as we can see in the example shown in Fig. 1, where we graphically represent two rules, $A \Rightarrow B$ and $A \Rightarrow C$. For the $A \Rightarrow B$ rule, we have the following support values for the intervening itemsets: supp $(A) = 28\%$, supp $(B) = 38\%$, and supp $(A \cup B) = 21\%$. Therefore, the confidence of the $A \Rightarrow B$ rule is 75%. For the $A \Rightarrow C$ rule, even though the support of the consequent changes (supp$(C) = 85\%$), the confidence value of the $A \Rightarrow C$ rule is still the same, 75%. In the first case, $B$ was present in 38% of the transactions in the database and its presence increases to 75% in transactions where $A$ is also present. In the second case, the presence of $A$ reduces the presence of $C$, from 85% to 75%. However, the confidence measure does not let us distinguish between these two different situations.

In short, the confidence measure does not take into account the support of the rule consequent, hence it is not able to detect negative dependencies between items. Several measures have been proposed in the literature as alternative interestingness measures in order to avoid the limitations of the traditional support/confidence framework [9]. In the following paragraphs, we describe some of them.

**Definition 4.** *The lift of the rule $A \Rightarrow C$, also known as interest, is defined as [4]:*

$$\textit{lift}\ (A \Rightarrow C) = \frac{\textit{supp}\ (A \Rightarrow C)}{\textit{supp}\ (A)\textit{supp}\ (C)}$$

The lift measure indicates how many times more often $A$ and $B$ occur together than expected if they where statistically independent. Values above 1 indicate positive dependence, while those below 1 indicate negative dependence. The lift values of the $A \Rightarrow B$ and $A \Rightarrow C$ rules in the aforementioned example are *lift* $(A \Rightarrow B) = 4.2$ and *lift* $(A \Rightarrow C) = 0.91$. Here, *lift* $(A \Rightarrow B) >$ *lift* $(A \Rightarrow C)$, which corresponds to our intuition that $A \Rightarrow B$ is more interesting than $A \Rightarrow C$.

Lift measures the degree of dependence between the itemsets. However, it only measures co-occurrence, but not the implication direction, since it is a symmetric measure, i.e., *lift* $(A \Rightarrow C) =$ *lift* $(C \Rightarrow A)$.

**Definition 5.** *The conviction of the rule $A \Rightarrow C$ is defined as [5]:*

$$\textit{conv}\ (A \Rightarrow C) = \frac{\textit{supp}(A)\textit{supp}(\neg C)}{\textit{supp}\ (A \cup \neg C)}$$

The advantage of conviction with respect to the confidence measure is that it takes into account both the support of the antecedent and the support of the consequent of the rule. Conviction values in the (0,1) interval mean negative dependence, values above 1 mean positive dependence, and a value of 1 means independence, as happened with the lift measure. Unlike lift, conviction is not a symmetric measure, i.e. it measures the implication direction.

In the example from Fig. 1, *supp* $(\neg B) = 0.62$ and *supp* $(A \cup \neg B) = 0.07$. Therefore, the conviction of the $A \Rightarrow B$ rule is 2.48. For the $A \Rightarrow C$ rule, *supp* $(\neg C) = 0.15$ and *supp* $(A \cup \neg C) = 0.07$. Therefore, $conv(A \Rightarrow C) = 0.6$, which means negative dependence.


Fig. 2. Graphical examples illustrating the gain (and the certainty factor) of the rules derived from the scenarios represented in Fig. 1: $A \Rightarrow B$ (left) and $A \Rightarrow C$ (right).

The main drawback of the conviction measure is that, as happened with lift, it is not bounded, i.e., its range is [0, $\infty$). Therefore, it is difficult to establish a convenient conviction threshold in practice.

Let us now define an ancillary measure that will be useful in our discussion below: the gain of a rule as the difference between its confidence and the support of its consequent. Formally,

**Definition 6.** *The gain of a rule $A \Rightarrow C$ is defined as:*

$$\mathit{gain}\ (A \Rightarrow C) = \mathit{conf}\,(A \Rightarrow C) - \mathit{supp}\ (C).$$

The gain values for the rules in Fig. 1 are $\mathit{gain}\ (A \Rightarrow B) = 0.75 - 0.38 = 0.37$ and $gain\ (A \Rightarrow C) = 0.75 - 0.85 = -0.10$. Figure 2 graphically shows these values. The lengths of the arrows represent the gain of the rules, i.e., the increase ($A \Rightarrow B$) or decrease ($A \Rightarrow C$) in the presence of the consequent given that the antecedent is present.

**Definition 7.** *The certainty factor of a rule $A \Rightarrow C$ is defined as [19]:*

$$\mathit{CF}\ (A \Rightarrow C) = \frac{\mathit{gain}\ (A \Rightarrow C)}{1 - \mathit{supp}\ (C)} \text{ if } \mathit{gain}\ (A \Rightarrow C) \geqslant 0, \text{ and}$$

$$\mathit{CF}\ (A \Rightarrow C) = \frac{\mathit{gain}\ (A \Rightarrow C)}{\mathit{supp}\ (C)} \text{ if } \mathit{gain}\ (A \Rightarrow C) < 0.$$

The certainty factor is, therefore, the gain value normalized into the [−1,1] interval. The certainty factor can be interpreted as a measure of the variation of the probability that the consequent is in a transaction when we consider only those transactions where the antecedent occurs. More specifically, a positive CF measures the increase of the probability that the consequent is in a transaction, given that the antecedent is.

In the example from Fig. 1, the CF of the $A \Rightarrow B$ rule is $0.37/(1 - 0.38) = 0.60$ while the CF for the $A \Rightarrow C$ rule is $-0.10/0.85 = -0.12$.

## 3. Interestingness measures for group association rules

In this section, we describe how association rules can be defined to study the features that individuals within a given group have in common. We define a *group* as a set of items $G = \{G_1, G_2, \ldots, G_n\}$ such that $G \subseteq I$. A *group association rule* $G : A \Rightarrow C$ is an association rule $A \Rightarrow C$ defined over the group $G$. In other words, a group association rule $G : A \Rightarrow C$ is equivalent to the classical association rule $GA \Rightarrow C$.

In the following paragraphs, we explain how to adapt the measures described in Section 2 to group association rules, as well as how these new measures can be useful for evaluating the interestingness of group association rules.

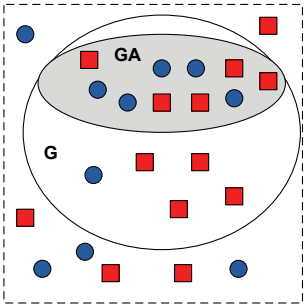



Fig. 3. Graphical representation of a group, $G$. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDA-130574)

### 3.1. Group support

**Definition 8.** *The support of an itemset $X$ in the group $G$ is the percentage of transactions in $G$ that contain $X$, i.e.,*

$$supp_G(X) = \frac{P(XG)}{P(G)}.$$

Figure 3 shows the representation of a group $G$ within an example dataset. The support of *circles (•)* in the group $G$ is $supp_G(\bullet) = 6/15 = 0.4$.

**Property 1.** *The support of an itemset $X$ in a group $G$ is the confidence of the rule $(G \Rightarrow X)$, i.e.,*

$$supp_G(X) = conf\,(G \Rightarrow X)$$

*Proof.* By Definition 8, $supp_G(X) = \frac{P(XG)}{P(G)}$. By Definition 3, $conf\,(G \Rightarrow X) = P(X|G) = \frac{P(XG)}{P(G)}$. Therefore, $supp_G(X) = conf\,(G \Rightarrow X)$ □

**Definition 9.** *The support of the group association rule $G : A \Rightarrow C$ is defined as:*

$$supp_G(A \Rightarrow C) = supp_G(A \cup C) = \frac{P(GAC)}{P(G)}$$

In the previous example, the support of the group association rule $G : A \Rightarrow \bullet$ is $supp_G(A \Rightarrow \bullet) = 5/15 = 0.33$.

**Property 2.** *The support of the rule $A \Rightarrow C$ in a group $G$ is the confidence of the rule $G \Rightarrow AC$, i.e.,*

$$supp_G(A \Rightarrow C) = conf\,(G \Rightarrow AC)$$

*Proof.* By Definition 9, $supp_G(A \Rightarrow C) = \frac{P(GAC)}{P(G)}$. By Definition 3, $conf\,(G \Rightarrow AC) = P(AC|G) = \frac{P(GAC)}{P(G)}$. Therefore, $supp_G(A \Rightarrow C) = conf\,(G \Rightarrow AC)$ □

### 3.2. Group confidence

**Definition 10.** *The confidence of the group association rule $G : A \Rightarrow C$ is defined as:*

$$conf_G(A \Rightarrow C) = \frac{supp_G(A \Rightarrow C)}{supp_G(A)}.$$

The confidence of the rule $G : A \Rightarrow \bullet$ in Fig. 3 is $conf_G(A \Rightarrow \bullet) = (5/15)/(10/15) = 0.5$.

**Property 3.** *The confidence of a rule $A \Rightarrow C$ in the group $G$ is the confidence of the rule $GA \Rightarrow C$ in the database, i.e.,*

$$conf_G(A \Rightarrow C) = conf\,(GA \Rightarrow C)$$

*Proof.* By Definition 10, $conf_G(A \Rightarrow C) = \frac{supp_G(A \Rightarrow C)}{supp_G(C)} = \frac{P(GAC)}{P(G)} \cdot \frac{P(G)}{P(GA)} = \frac{P(GAC)}{P(GA)}$. By Definition 3, $conf\,(GA \Rightarrow C) = \frac{P(GAC)}{P(GA)}$. Therefore, $conf_G(A \Rightarrow C) = conf\,(GA \Rightarrow C)$. □

### 3.3. Group gain

**Definition 11.** *The gain of the group association rule $G : A \Rightarrow C$ is defined as:*

$$gain_G(A \Rightarrow C) = conf_G(A \Rightarrow C) - supp_G(C)$$

The gain represents the difference between the confidence in the presence of the consequent when we know that the antecedent appears in the group, minus the support of the consequent within the group.

In Fig. 3, the support of the *circles* in the group G is $supp_G(\bullet) = 6/15 = 0.4$ and the confidence of the $G : A \Rightarrow \bullet$ rule is $conf_G(A \Rightarrow \bullet) = 0.5$. Then, the gain of the rule is $gain_G(A \Rightarrow \bullet) = 0.5 - 0.4 = 0.1$. That means that, within the group $G$, finding a *circle* is 10% more likely when $A$ holds.

Rules with high positive gain values help us describe subgroups (circles in the previous example) within the group $G$. On the other side, rules with high negative gain values help us find characteristics that are less frequent within the subgroup than in the overall group. For example, if the rule $A \Rightarrow \bullet$ had a negative gain value, that would mean that it would be more difficult to find a circle among those elements in $GA$ than in $G$.

**Property 4.** *The gain of the rule $G : A \Rightarrow C$ is the difference between the confidence of the $GA \Rightarrow C$ rule and the confidence of the $G \Rightarrow C$ rule , i.e.,*

$$gain_G(A \Rightarrow C) = conf\,(GA \Rightarrow C) - conf\,(G \Rightarrow C)$$

*Proof.* By Definition 11, $gain_G(A \Rightarrow C) = conf_G(A \Rightarrow C) - supp_G(C)$. By Properties 2 and 3, $supp_G(A \Rightarrow C) = conf\,(G \Rightarrow AC)$ and $conf_G(A \Rightarrow C) = conf\,(GA \Rightarrow C)$. Therefore, $gain_G(A \Rightarrow C) = conf\,(GA \Rightarrow C) - conf\,(G \Rightarrow C)$ □

**Property 5.** *The gain of the $G : A \Rightarrow C$ rule is the difference between the gain of the $GA \Rightarrow C$ rule and the gain of the $G \Rightarrow C$ rule , i.e.,*

$$gain_G(A \Rightarrow C) = gain\,(GA \Rightarrow C) - gain\,(G \Rightarrow C)$$

*Proof.* By Definition 6, $gain\ (G \Rightarrow C) = conf\,(G \Rightarrow C) - supp\ (C)$. Then, we can solve for $conf\,(G \Rightarrow C)$ as $conf\,(G \Rightarrow C) = gain\ (G \Rightarrow C) + supp\ (C)$. If we replace the $conf\,(G \Rightarrow C)$ in Property 4, we obtain $gain_G(A \Rightarrow C) = conf\,(GA \Rightarrow C) - conf\,(G \Rightarrow C) = conf\,(GA \Rightarrow C) - supp\ (C) - gain\ (G \Rightarrow C)$. Finally, by Definition 6, $conf\,(GA \Rightarrow C) - supp(C) = gain\ (GA \Rightarrow C)$. Therefore, $gain_G(A \Rightarrow C) = gain\ (GA \Rightarrow C) - gain\ (G \Rightarrow C)$. □

**Theorem 1. Gain pseudo-commutativity**. *The difference between the gain of the $G : A \Rightarrow C$ rule in the $G$ group and the gain of the $A \Rightarrow C$ rule in the database equals the gain of the rule $A : G \Rightarrow C$ in the group $A$ minus the gain of the $G \Rightarrow C$ rule in the database, i.e.,*

$$gain_G(A \Rightarrow C) - gain\ (A \Rightarrow C) = gain_A(G \Rightarrow C) - gain\ (G \Rightarrow C).$$

*Proof.* By Definition 6, $gain\ (G \Rightarrow C) = conf\,(G \Rightarrow C) - supp(C)$.

We can isolate $conf\,(G \Rightarrow C) = gain\ (G \Rightarrow C) + supp\ (C)$ and replace it in $gain_G(A \Rightarrow C) = conf\,(GA \Rightarrow C) - conf\,(G \Rightarrow C) = conf\,(GA \Rightarrow C) - (gain\ (G \Rightarrow C) + supp\ (C))$.

If we isolate $supp\ (C)$ from Definition 6 and replace it in the previous expression, we obtain: $gain_G(A \Rightarrow C) = conf\,(GA \Rightarrow C) - (gain\ (G \Rightarrow C) + supp\ (C)) = conf\,(GA \Rightarrow C) - gain\ (G \Rightarrow C) - (conf\ (A \Rightarrow C) - gain\ (A \Rightarrow C)) = conf\,(GA \Rightarrow C) - (conf\ (A \Rightarrow C) - gain\ (G \Rightarrow C) + gain\ (A \Rightarrow C))$.

By Definition 11, the first term can be expressed as $conf\,(GA \Rightarrow C) - conf\,(A \Rightarrow C) = gain_A(G \Rightarrow C)$. Then, we have $gain_G(A \Rightarrow C) = gain_A(G \Rightarrow C) - gain\ (G \Rightarrow C) + gain\ (A \Rightarrow C))$.

Therefore, $gain_G(A \Rightarrow C) - gain\ (A \Rightarrow C) = gain_A(G \Rightarrow C) - gain\ (G \Rightarrow C)$. □

### *3.4. Group gain normalization*

The range of the gain measure for group association rules is $[-supp_G(C),\ 1 - supp_G(C)]$. In the following paragraphs, we propose several ways to normalize the group gain depending on the kind of information the user might be more interested in.

#### *3.4.1. Group gain factor*

We can normalize the gain into the $[-1, 1]$ interval to obtain a gain factor measure that corresponds to the certainty factor in the general association rule framework:

**Definition 12.** *The gain factor of the group association rule $G : A \Rightarrow C$ is defined as:*

$$GF_G(A \Rightarrow C) = \frac{gain_G(A \Rightarrow C)}{1 - supp_G(C)} \text{ if } gain_G(A \Rightarrow C) \geqslant 0, \text{ and}$$

$$GF_G(A \Rightarrow C) = \frac{gain_G(A \Rightarrow C)}{supp_G(C)} \text{ if } gain_G(A \Rightarrow C) < 0.$$

In our example, the gain factor of the rule $G : A \Rightarrow \bullet$ is $GF_G(A \Rightarrow \bullet) = 0.1/(1 - 0.4) = 0.17$.

This measure is proportional to the group gain. When it is positive, it is also inversely proportional to the value $[1 - supp_G(C)]$. Therefore, all other things being equal, $GF$ will be larger for subgroups of elements that were more common in the group G (i.e., those having a higher $supp_G(C)$). When GF is negative, it is inversely proportional to $supp_G(C)$: it will have a larger absolute value when the subgroup is less frequent in $G$, i.e. for small values of $supp_G(C)$.

*3.4.2. Group variation*

The group variation measure always normalizes the gain using the $supp_G(C)$ value, in order to highlight those consequents that are less frequent in our database.

**Definition 13.** *The variation of a group association rule $G : A \Rightarrow C$ is defined as:*

$$\delta_G(A \Rightarrow C) = \frac{gain_G(A \Rightarrow C)}{supp_G(C)} = \frac{conf_G(A \Rightarrow C) - supp_G(C)}{supp_G(C)}$$

In contrast to the GF, variation is inversely proportional to $supp_G(C)$ when it is positive. It will have a higher value the less frequent $C$ is in $G$. It should be noted that the variation equals the gain factor when the gain is negative. The variation of the rule $G : A \Rightarrow \bullet$ in Fig. 3 is $\delta_G(A \Rightarrow \bullet) = 0.1/(0.4) = 0.25$.

*3.5. Group impact*

In this section we present two new interestingness measures that are also based on the group gain. In these measures, we take into account the support of the group corresponding to the antecedent of the rule.

*3.5.1. Impact*

**Definition 14.** *The impact of the group association rule $G : A \Rightarrow C$ is defined as:*

$$impact_G(A \Rightarrow C) = supp(GA) * gain_G(A \Rightarrow C)$$

The impact of a group association rule represents the number of individuals that are affected by the rule, i.e., the number of individuals that we did not expect to find in the transactions within the group G that contain $A$ given what we knew about $G$ in general.

The impact is proportional to $gain_G(A \Rightarrow C)$ and $supp\ (GA)$. It will be higher for those rules with a high gain and a frequent antecedent in the $G$ group.

In our example from Fig. 3, $impact_G(A \Rightarrow \bullet) = (10)*0.1 = 1$. That impact means that there is one circle that we did not expect to find in $GA$ when we only knew the support of circles in $G$. Since $supp_G(\bullet) = 0.4$, we did expect four circles in $GA$ but found five of them.

*3.5.2. Impact ratio*

The impact measure represents the number of individuals that are affected by the rule. However, in most cases, groups have different numbers of individuals and an absolute impact value might be misleading. The impact ratio might be useful in such situations, since it represents the ratio between the number of individual affected by the rule and the number of individuals within the group.

**Definition 15.** *The impact ratio of the group association rule $G : A \Rightarrow C$ is defined as:*

$$IR_G(A \Rightarrow C) = \frac{impact_G(A \Rightarrow C)}{supp\ (G)}$$

The impact ratio of a group association rule represents the proportion of the impact of the rule within the group $G$ with respect to the size of the group. The impact ratio is $IR_G(A \Rightarrow \bullet) = 1/15 = 0.07$ in the example from Fig. 3.

Table 1
Values of the different interestingness measures for the $(A \Rightarrow \bullet)$ and $(B \Rightarrow \blacksquare)$ rules in Fig. 4

| | $A \Rightarrow \bullet$ | $B \Rightarrow \blacksquare$ |
|---|---|---|
| $supp_G(Y)$ | $4/20 = 0.20$ | $14/20 = 0.70$ |
| $supp_G(X \Rightarrow Y)$ | $3/20 = 0.15$ | $4/20 = 0.20$ |
| $conf_G(X \Rightarrow Y)$ | $3/10 = 0.30$ | $4/5 = 0.80$ |
| $gain_G(X \Rightarrow Y)$ | $0.3 - 0.2 = 0.10$ | $0.8 - 0.7 = 0.10$ |
| $GF_G(X \Rightarrow Y)$ | $0.1/0.8 = 0.125$ | $0.1/0.3 = 0.33$ |
| $\delta_G(X \Rightarrow Y)$ | $0.1/0.2 = 0.50$ | $0.1/0.7 = 0.14$ |
| $impact_G(X \Rightarrow Y)$ | $0.1 * 10 = 10$ | $0.1 * 5 = 0.50$ |
| $IR_G(X \Rightarrow Y)$ | $1/20 = 0.05$ | $0.5/20 = 0.025$ |

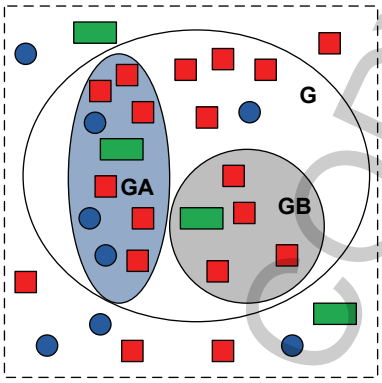



Fig. 4. Graphical representation of a group, G, and two rules within that group, $(A \Rightarrow \bullet)$ and $(B \Rightarrow \blacksquare)$. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDA-130574)



## 4. Ranking groups and group association rules

All the different groups that can be identified in the database, as well as the rules within them, can be obtained in an unsupervised manner. However, the number of rules and groups obtained in the rule mining process can be huge. As a consequence, a second-order data mining problem arises: it may be too difficult to extract useful information from so many rules and groups.

In this section, we explain how to rank the groups (and the rules within the groups) according to their potential interestingness. We will use the measures we described in previous sections to highlight those rules and groups that might be relevant to the user according to different criteria.

### *4.1. Ranking rules within a particular group*

The use of each different interestingness measure provides a different ordering relationship among the discovered association rules. In this section, we analyze how two rules within a group will have a different relative order within that group depending on the interestingness measure we use to evaluate them.

Figure 4 shows a group, $G$, within an example dataset. In this group, we have identified two rules, $A \Rightarrow \bullet$ and $B \Rightarrow \blacksquare$, whose values for the different interestingness measures we described in the previous sections are summarized in Table 1.

#### 4.1.1. Characterizing subgroups within a group

If we are interested in obtaining those rules that characterize subgroups within a group (i.e., rules sharing their consequent), we should use the *gain* measure because a high gain increases our confidence in the presence of the consequent when we know that the antecedent holds.

In our example from Fig. 4, the gain of both rules is 10%, i.e., the confidence on the presence of • increases if $A$ also holds, while the confidence on the presence of the ■ increases, in the same amount, when B is true.

Both are important rules within the group $G$ but they give us different information as squares are more frequent in $G$ than circles. Therefore, we should use other measures for distinguishing between them:

- If we want to highlight the most frequent subgroups, the *gain factor*, as inversely proportional to the interval $[1 - supp_G(C)]$, should be used.
  In our example, the $GF_G$ value is higher for the $B \Rightarrow$ ■ rule because squares where more frequent in $G$ than circles.
- If we want to highlight anomalies, the *variation* measure is a better choice since, in contrast to the gain factor, it overweighs those subgroups that have a low support in the group. That is the case of the circles in G: the $A \Rightarrow$ • rule has a larger value of $\delta_G$.

#### 4.1.2. Characterizing subgroups using frequent itemsets

If we are interested, not only in the subgroups themselves, but also in discovering itemsets that make them distinct in our database, we should use an interestingness measure that takes into account the frequency of the rule antecedent, e.g., the *impact* measure.

In our example, the impact of the $A \Rightarrow$ • rule in G is larger than the impact of the $B \Rightarrow$ ■ rule because the support of A is larger than the support of B.

This measure has the advantage that it can be easily interpreted: the number of individuals in $G$ that are directly affected by the $A \Rightarrow C$ rule, i.e., those individuals that are not expected to be in $GA$ given the overall support of $C$ in the group.

In our example, the impact of the $A \Rightarrow$ • rule is 1 because, given a 20% support for circles in $G$ and 10 elements in the $GA$ subgroup, we expected 2 circles in $GA$ but found 3 of them.

It should be noted that the impact ratio measure gives us the same relative ordering among rules within the same group than the impact measure, since it takes into account the size of the group $G$, which is the same for all the rules within the same group.

### 4.2. Ranking groups within the database

Apart from ordering rules within a group, we can establish an order relationship among the groups in our database to highlight those groups that host the most interesting rules.

Many different association rules can hold within a given group, but not all of them are equally important to describe the group. When we intend to rank groups rather than individual rules, we must somehow compute an aggregate value that represents the overall interestingness of the group. We should average the impact of the $n$ rules $\{(A_i \Rightarrow C_i)\}_{i=1\cdots n}$ within a given group to indicate the overall interestingness of that group. However, as we have explained in Section 4.1, some rules are more interesting than others, hence they should not have the same importance when computing the overall group score.

Impact seems to be a good candidate for estimating the potential interestingness of the group, since it takes into account the number of individuals that are affected by each rule within the group. We can

then define the weighted impact for the rules within a group using different interestingness measures. Formally, we define the weighted impact as:

$$\text{Weighted impact } (G) = \frac{\sum_{i=1}^{n} I_G(A_i \Rightarrow C_i) \cdot \mathit{impact}_G(A_i \Rightarrow C_i)}{\sum_{i=1}^{n} I_G(A_i \Rightarrow C_i)}$$

where $I_G(A_i \Rightarrow C_i)$ represents the value of any of the interestingness measures described in Section 3 and analyzed in Section 4.1, for each $A_i \Rightarrow C_i$ rule in the $G$ group.

Large groups tend to have higher impact values for their rules because their impact depends on the support of their antecedent within the group and it is usually larger in large groups. Therefore, small groups are penalized in the overall group ranking if we use the impact measure to average the interestingness of the rules within the group. When we want to take into account the relative size of the groups, we should use the impact ratio measure instead, which gives us a more balanced ranking for groups of disparate size. Thus, we can also define a weighted impact ratio measure to rank groups within the database:

$$\text{Weighted IR } (G) = \frac{\sum_{i=1}^{n} I_G(A_i \Rightarrow C_i) \cdot IR_G(A_i \Rightarrow C_i)}{\sum_{i=1}^{n} I_G(A_i \Rightarrow C_i)}$$

## 5. Comparing alternative ranking criteria

Once we have proposed different interestingness measures that can be used to rank groups and group association rules, we are interested in comparing the rankings obtained by each measure in order to analyze how similar (or different) they are.

Several measures have been proposed in the literature to compare two permutations $\sigma_1$ and $\sigma_2$ whose elements are in $D$. Two well-known measures are [7]:

- *Kendall's tau:* For each pair $i, j \in P$ of distinct members of $D$, if $i$ and $j$ are in the same order in $\sigma_1$ and $\sigma_2$, then let $K_{i,j}(\sigma_1,\sigma_2) = 0$; and if $i$ and $j$ are in the opposite order, then $K_{i,j}(\sigma_1,\sigma_2) = 1$. Kendall's tau is given by $K(\sigma_1,\sigma_2) = \sum_{\{i,j\} \in P} K_{i,j}(\sigma_1, \sigma_2)$. The maximum value of $K(\sigma_1, \sigma_2)$ is $n(n-1)/2$, where $n$ is the number of elements in $\sigma_j$. This maximum occurs when $\sigma_1$ is the reverse of $\sigma_2$; that is, when $\sigma_1(i) + \sigma_2(i) = n+1$ for each $i$, being $\sigma_j(i)$ the index of the $i$-th element in the $\sigma_j$ permutation. Kendall's tau turns out to be equal to the number of exchanges needed in a bubble sort to convert one permutation into the other. Kendall's tau can be computed in $O(nlogn)$ using a divide-and-conquer algorithm [13].
- *Spearman's footrule* is defined by $F\ (\sigma_1,\ \sigma_2) = \sum_{i=1}^{n} |\sigma_1(i) - \sigma_2(i)|$. The maximum value of $F(\sigma_1,\sigma_2)$ is $n^2/2$ when $n$ is even, and $(n+1)(n-1)/2$ when n is odd. As with Kendall's tau, the maximum occurs when $\sigma_1$ is the reverse of $\sigma_2$. This measure can be easily computed in $O(n)$ given an inverse index for both rankings.

These measures let us compare two complete rankings. However, in many cases, the top K elements in the rankings are more important than others that appear at less prominent positions (think of a Web search engine, for instance). Fagin [7] proposed an adaptation of the aforementioned measures to compare two top-K lists, $\tau_1$ and $\tau_2$:

- *Kendall's tau for top-K lists*: We have to consider 4 possible scenarios taking into account that not all the elements in the top K list $\tau_1$ may be present in the top K list $\tau_2$ and vice versa. For each pair $i, j \in D_{\tau_1} \cup D_{\tau_2}$:

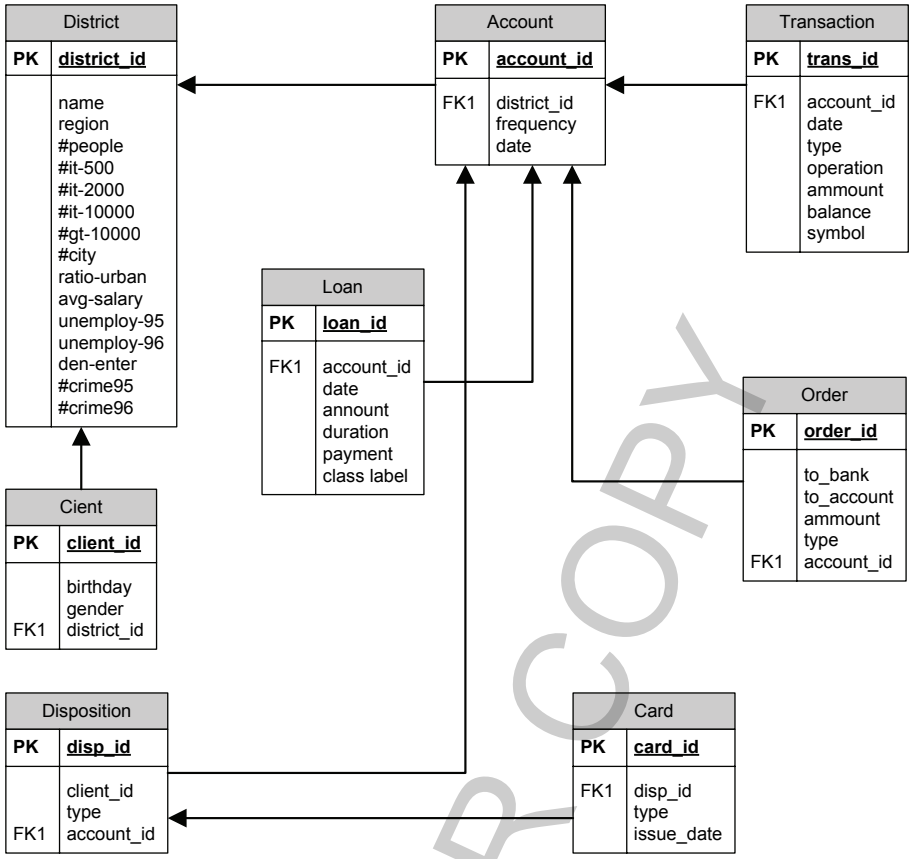



Fig. 5. Loan database schema.

- Case 1: $i, j$ are both present in $\tau_1$ and $\tau_2$. Then, if they are in the same order in both lists, $K_{i,j}(\tau_1, \tau_2) = 0$ while, if they are in the opposite order, $K_{i,j}(\tau_1,\tau_2) = 1$.
- Case 2: $i, j$ are both present in $\tau_1$ but only $i$ is present in $\tau_2$. Then, if $i$ is ahead of $j$ in $\tau_1$, $K_{i,j}(\tau_1,\tau_2) = 0$. Otherwise, $K_{i,j}(\tau_1,\tau_2) = 1$.
- Case 3: $i$ is only present in $\tau_1$ and $j$ is only present in $\tau_2$. Then, $K_{i,j}(\tau_1,\tau_2) = 1$.
- Case 4: $i, j$ are both present in one top K list (say $\tau_1$) but neither $i$ nor $j$ appear in the other top K list (say $\tau_2$). Then, we do not have information about the order of $i$ and $j$ in $\tau_2$. Therefore, $K_{i,j}(\tau_1,\tau_2) = p$, where the value of $p$ is usually 0.5 to be neutral.

– The *Spearman's footrule for top-K lists* is computed as $F^l(\tau_1,\tau_2) = \sum_{i \in D_{\tau_1} \cup D_{\tau_2}} |\tau'_1(i) - \tau'_2(i)|$, where $x \in \{1, 2\}$, and $\tau'_x(i)$=$\tau_x(i)$ for $i \in D_{\tau_x}$, otherwise $\tau'_x(i) = l$. A natural choice for $l$ is K+1.

## 6. Experimental results

In order to study the performance of the proposed interestingness measures, we have performed a series of experiments using the `loan` multirelational database. The `loan` database was used in the PKDD CUP'09. Figure 5 shows the schema of this database.

The database contains eight relations with 75,982 tuples in total. The analysis of a multirelational database typically starts from a particular relation. This relation, which we will call *target relation*, is

selected by the end user according to his specific goals. In our case, the target relation is `account` which contain 4500 tuples. This database was adapted by Yin et al. for their experiments with CrossMine [24] and can be downloaded from: http://research.microsoft.com/en-us/people/xyin/.

### 6.1. Using trees to mine multirelational databases

A common approach to mine multirelational databases consists in joining all the relations in the database in order to obtain a single relation, usually called universal relation [8,15,16]. Then, classical data mining techniques can be applied to this universal relation. However, join-based techniques such as the aforementioned one present a serious disadvantage: they do not preserve the proper support counts.

We have proposed two alternative representation schemes for multirelational databases. Our representation schemes are based on trees, so that we can apply existing tree mining techniques to identify frequent patterns in multirelational databases [11].

The main idea behind our two representation schemes is building a tree from each tuple in the target relation and following the links between relations (i.e. foreign keys) to collect all the information related to each tuple in the target relation.

The key-based tree representation scheme is inspired by the concept of identity in the relational model while the object-based one is based on the concept of identity in object-oriented models. Relational databases rely on primary keys to ensure that each table row can be univocally referenced. Any unique field, or combination of fields, can be used as the primary key. In the object model, however, each object is already unique, and no specific key is needed. In an object database, each object is automatically assigned an unique ID. This does mean that you can create objects that have identical field values but are still different objects [17].

In our experiments, we have considered the `account` relation as our target relation in the `loan` database and we have built trees by collecting information from the `loan`, `disposition`, `order`, and `district` relations.

The tree database obtained from the `loan` multirelational database contains 4500 trees. Trees have an average of 34 nodes using the key-based representation scheme (153,102 nodes in total) and an average of 37 nodes using the object-based one (169,842 nodes in total).

### 6.2. Extracting group association rules from multirelational databases

We have transformed the `loan` database into two sets of trees (using both the key-based and object-based tree representation schemes). This way, tree mining algorithms can be used to extract patterns from this kind of databases, as described above. We can also define two different kinds of patterns to be identified within the trees: induced and embedded patterns. Therefore, four different kinds of patterns can be mined from a multirelational database using our approach, i.e., induced key-based patterns, embedded key-based patterns, induced object-based patterns, and embedded object-based patterns.

We use in [12] a tree pattern mining algorithm called POTMiner to identify these four different kinds of patterns within the `loan` database. Once we have identified the frequent tree patterns derived from the multirelational database, we have extracted association rules from them, using techniques that are analogous to the ones employed in a more traditional setting. Group association rules can then be derived from the discovered association rules just by clustering rules sharing parts of their antecedents.

Figure 6 shows the number of identified groups within the `loan` database for the four different kinds of patterns (induced key-based patterns, embedded key-based patterns, induced object-based patterns,

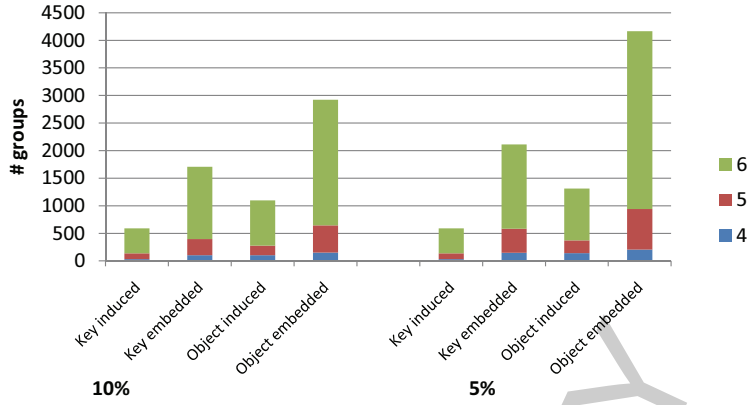



Fig. 6. Number of identified groups within the `loan` database, using patterns with varying size (from 4 to 6) and two different support thresholds (10% and 5%). (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDA-130574)

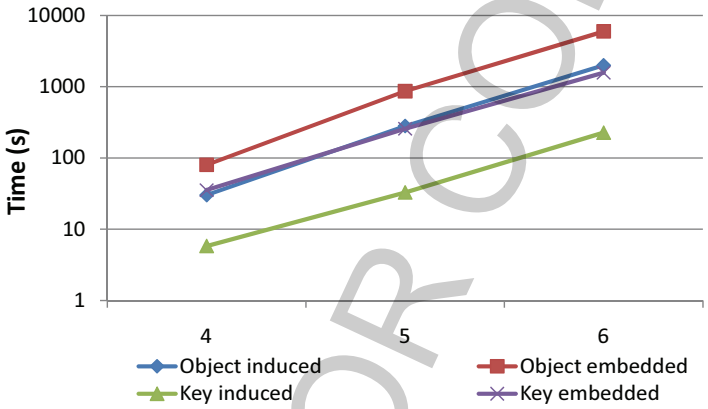


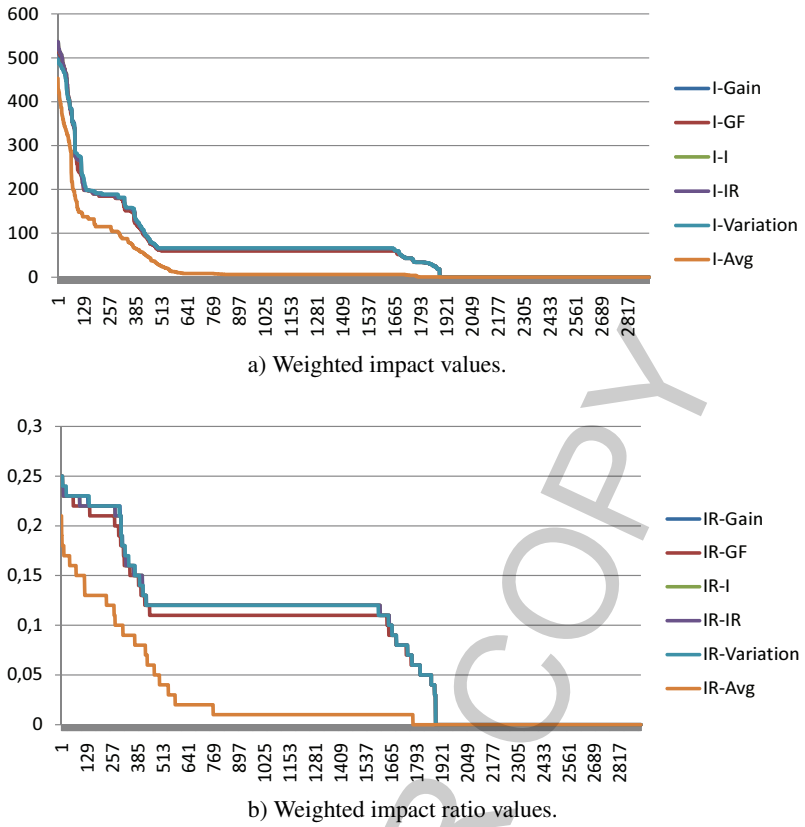

Fig. 7. POTMiner execution time required to extract groups from the `loan` multirelational database from frequent patterns of increasing size. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDA-130574)

and embedded object-based patterns) and two different minimum support thresholds (10% and 5%). As expected, the number of identified groups increases when the size of the patterns also increases, and also when the minimum support threshold decreases, since the more patterns are identified, the more group association rules can be extracted.

Figure 7 shows the execution time required by our algorithm to obtain all the existing groups in the multirelational database, where time is displayed in seconds on a logarithmic scale. It should be noted that execution time includes the whole process required for identifying group association rules: the time required to identify the patterns, the time required to identify the association rules, and the time required to group the association rules into sets of group association rules. The identification of key-induced patterns is the fastest because there is a lower number of this kind of patterns. On the other hand, mining object-embedded patterns is the most costly alternative because more patterns are involved.

### 6.3. Ranking groups

In this section, we compare the group rankings provided by different interestingness measures in order to check how many groups are highlighted as the most interesting ones and analyze potential differences




a) Weighted impact values.

b) Weighted impact ratio values.

Fig. 8. Weighted impact and weighted impact ratio values for the 2924 groups identified within the `loan` multirelational database using six different interestingness measures. (Colours are visible in the online version of the article; http://dx.doi.org/10.3233/IDA-130574)

among them.

In these series of experiments, we have used the groups and rules obtained from the `loan` multirelational database, using embedded object-based patterns up to size 6 (i.e. maxsize = 6) and a 10% minimum support threshold. Using these parameters, we obtained 2,924 groups and 7,394,860 group association rules.

We have ranked the resulting groups using the weighted impact and the weighted impact ratio measures. In order to weight those measures, we have tested five different interestingness measures: gain, gain factor, impact, impact ratio, and variation. Therefore, we have obtained 10 alternative rankings for the groups in our database. We have also included two additional rankings using the average impact and the average impact ratio measures, just to check that weighted measures give us different results than their plain averages (i.e., just to corroborate that taking rule interestingness into account is preferred over just aggregating rule impacts without taking that information into account).

Figures 8 a) and b) show the values of the weighted impact and the weighted impact ratio measures, respectively, for every group within the `loan` multirelational database (sorting the groups according to their weighted evaluation measure).

As it can be seen in Fig. 8 a), the weighted impact measures highlight about 130 groups in the database. It should be noted that the weighted impact provides similar results for every different interestingness measure, i.e. the rankings provided by the different interestingness measures, used as weights in the

weighted impact formula, are quite similar. The average values are lower, as expected, since less interesting rules contribute less to diminish the overall group score when using the weighted measure. The similarities among the particular rankings provided by different interestingness measures will be discussed later.

Figure 8 b) shows that the weighted impact ratio measures highlight about 400 of the groups in the database, more than the weighted impact measures. The weighted impact tends to highlight only large groups, since the impact of the rules depends on the absolute support of the antecedent in the group association rules and that support is typically larger in large groups. Using the weighted impact ratio, however, the group size does not influence the final group score because impact ratio is a relative measure. Hence the higher number of highlighted groups when using the weighted impact ratio. This way, smaller groups are highlighted when interesting rules affect a significant portion of them.

As happened with the impact measures, the results using different interestingness measures as weights for the weighted impact ratio are similar and raw averages provide lower scores. In the following section, we analyze the particular rankings provided by each alternative evaluation criterion.

### *6.4. Comparing rankings provided by different interestingness measures*

In the previous section, we obtained twelve different rankings of the 2,924 groups within the `loan` database. In this section, we use the Kendall's and Spearman's measures we described in Section 5 to compare these rankings in order to discover their similarities.

We have compared every pair of rankings using both similarity measures. Figure 9 shows the resulting similarity matrices using Kendall's (top) and the Spearman's (bottom) measures. Black cells indicate that the rankings are almost indistinguishable (their Kendall or Spearman value is almost 0), while white cells correspond to the most dissimilar pairs of rankings within the matrix (i.e., they have the maximum Kendall or Spearman value within all the pairs).

As it can be seen in Fig. 9, we can easily observe two clearly-defined groups of different rankings, which correspond to those obtained using the weighted impact measures (upper left quadrant) and the ones obtained using the weighted impact ratio measures (lower right quadrant). When interpreting the results, it is important to recall that Kendall's measure is an order-based measure while Spearman's measure is a distance-based one.

In Fig. 9, we are comparing the rankings for all the 2,924 groups within our database, hence there are some differences in the comparison between different variants of the same kind, a fact that is clearly visible for the weighted impact ratio results (lower right quadrant), especially for Kendall's order-based measure. Spearman's distance-based measure highlights some differences when using the gain factor as interestingness measure for individual rules within a group, as well as a surprising similarity between the average impact and the average impact ratio rankings.

However, it should be noted that apparent differences do not have the same importance if they happen at the top of the rankings or at their bottom, since end users will not usually delve into the complete list of 2,924 groups. Therefore, a more realistic scenario consists in comparing the top elements in each ranking.

We have compared the top 100 and the top 10 groups in the rankings provided by each ranking criterion in order to obtain a more realistic comparison of ranking results.

The matrices shown in Fig. 10 show the values for Kendall's and Spearman's measures when comparing the top 100 groups in each of the rankings. In this case, the matrices we have obtained using both Kendall's order-based and Spearman's distance-based similarity measures are remarkably similar. As

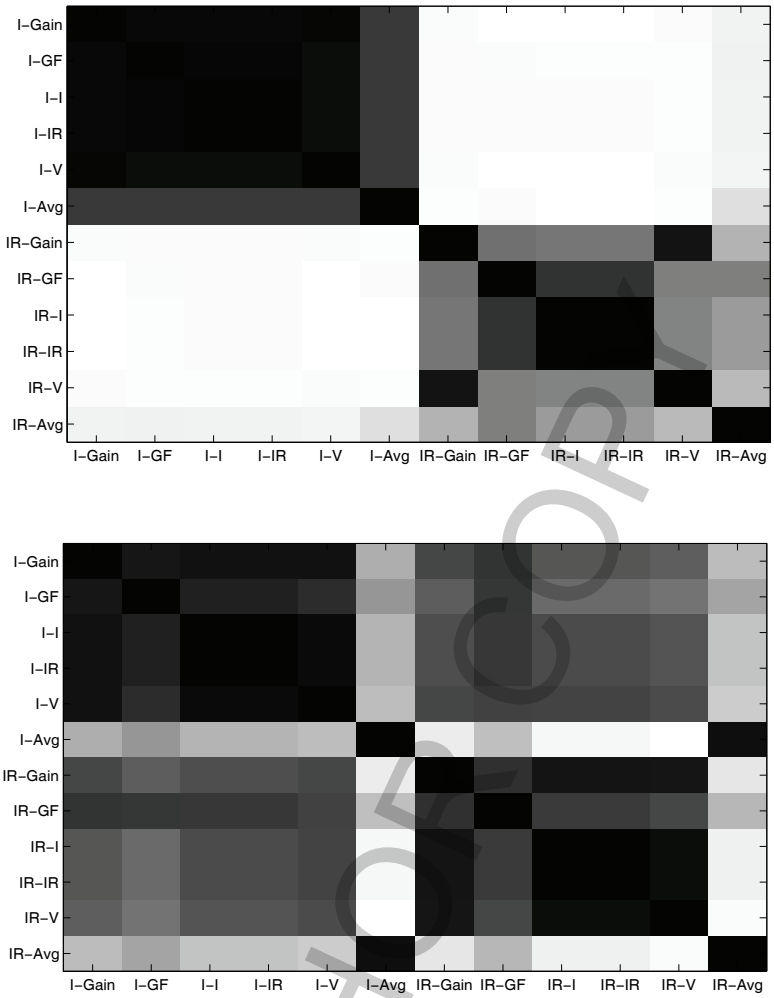



Fig. 9. Similarity matrices using Kendall's (top) and the Spearman's (bottom) measures to compare the rankings.

happened when comparing the whole rankings, two groups are clearly visible: the ones corresponding to the weighted impact and those resulting from using the weighted impact ratio.

The different variants of weighted impact are almost undistinguishable, while some differences exist among the different rankings provided by the weighted impact ratio. Within the latter, using the impact (IR-I) and the impact ratio (IR-IR) as weights provide very similar results. Likewise, gain (IR-Gain) and variation (IR-V) are also similar. However, there are differences between gain factor (IR-GF) and variation (IR-V). Gain factor equals variation, but only for negative gain values. Here, we are considering the 100 most interesting groups, which typically contain many rules with positive gain values (it should be recalled that both impact and impact ratio are proportional to gain values).

Finally, we have compared the top 10 groups in the rankings as shown in Fig. 11. For our particular database, we can again identify two clusters around weighted impact and weighted impact ratio measures. There are still some differences between the weighted measures and their plain averages, as well

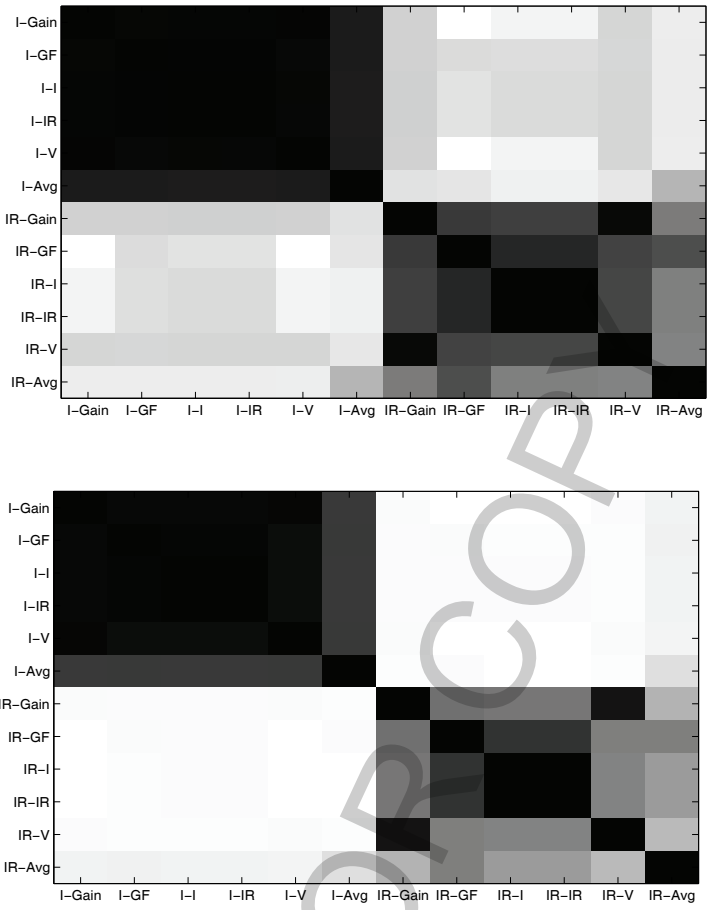



Fig. 10. Similarity matrices using Kendall's (top) and Spearman's (bottom) measures to compare the top 100 groups in each ranking.

as between the variation variant of the weighted impact (I-V) and the rest of its variants (I-Gain, I-GF, I-I, I-IR). In this case, the I-V results are more similar to the I-Gain and the I-GF results (whose interestingness measures are closer to variation, which is just a method for normalizing rule gain, as the gain factor) than they are to the impact-related measures (I-I and I-IR).

## 7. Identifying the most interesting groups

In the previous section we have obtained 4 different sets of rankings for the groups in our experiments:

1. The one obtained from the average impact of the rules in each group.
2. The one obtained from the average ratio of the rules in each group.
3. The one obtained using the weighted impact measure (using any interestingness measure as weight).
4. The one obtained using the weighted impact ratio measure (using any interestingness measure as weight).

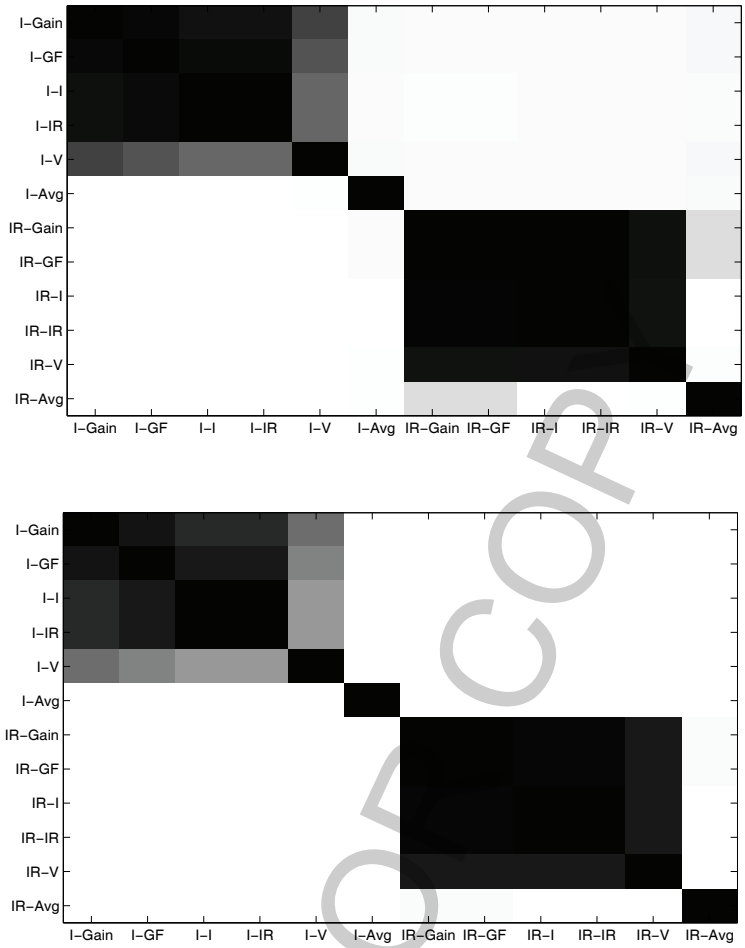



Fig. 11. Similarity matrices using Kendall's (top) and Spearman's (bottom) measures to compare the top 10 groups in each ranking.

When using the average impact of the rules within each group (I-Avg), large groups are favored. Almost all the groups in the top 10 have a 100% support and most of them are trivial and not really interesting. The average impact ratio (IR-Avg) is not so biased towards large groups.

Using the weighted impact measure (I-Gain, I-GF, I-I, I-IR, and I-V) provides a more balanced ranking. Albeit the support of the groups in their top 10 is still high (around 75%), end users can find some interesting groups, such as "accounts that have a monthly frequency of issuance of statements and an owner-type disposition".

The weighted impact ratio measure (IR-Gain, IR-GF, IR-I, IR-IR, IR-V) provides more interesting results. In this case, the support of the groups in the top 10 is around 25% and much more specific groups are highlighted, such as those "accounts that have a monthly frequency of issuance of statements, and a house-type permanent order, and are located in a district with one municipality with more than 10000 inhabitants".

In summary, weighted impact is biased towards large groups and it should be used when looking for the characterization of very large clusters within a database of individuals. However, users should resort to the weighted impact ratio if they are willing to discover smaller groups of potentially more interesting individuals (i.e. those who are not so common in the database but exhibit consistently different peculiar behaviors).

## 8. Conclusions

Databases naturally contain groups of individuals that share some of their features and some aspects of their behavior. In this paper, we have proposed group association rules, which are association rules that can be discovered within groups of individuals.

We have adapted some of the standard interestingness measures for association rules to group association rules. We have also proposed new interestingness measures to evaluate group association rules and we have studied some of the formal properties of such measures.

Finally, we have proposed an approach to rank groups within a database (and also rules within each group). Alternative rankings can be provided by employing alternative, and often complementary, interestingness measures. Depending on their particular goals, users should choose which measure to employ in order to highlight the most potentially interesting groups. When the user's interest is to highlight large groups, then he should choose a weight impact measure. On the other hand, if he is more interested in discovering small groups of individuals with common peculiar behavior, then the weighted impact ratio is the appropriate measure.

Our experiments on a real-world database corroborated our intuitions on the behavior of different ranking criteria and demonstrated that our approach can be useful in practice for ameliorating the second-order data mining problem that users must face when dealing with the huge amount of association rules that can be derived from real-world databases (more than seven million in our case study).



## Acknowledgements

Work partially supported by the TIN2009-08296 and TIN2012-36951 research projects funded by the Spanish Ministry of Science and Innovation.

We are grateful to the anonymous referees for their valuable comments and suggestions.

## References

[1] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, in: *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994, pp. 487–499.

[2] F. Berzal, I.J. Blanco, D. Snchez and M.A. Vila, Measuring the accuracy and interest of association rules: A new framework, *Intelligence Data Analysis* **6**(3) (2002), 221–235.

[3] FãĂĆ Berzal and J.C. Cubero, Guest editors' introduction, *Data and Knowledge Engineering* **60**(1) (2007), 1–4.

[4] S. Brin, R. Motwani and C. Silverstein, Beyond market baskets: Generalizing association rules to correlations, in: *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, volume 26, 1997, pp. 265–276.

[5] S. Brin, R.Motwani, J.D. Ullman and S. Tsur, Dynamic itemset counting and implication rules for market basket data, in: *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, volume 26, 1997, pp. 255–264.

[6] W. DuMouchel and D. Pregibon, Empirical bayes screening for multi-item associations, in: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, 2001, pp. 67–76.

[7] R. Fagin, R. Kumar and D. Sivakumar, Comparing top k lists, in: *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2003, pp. 28–36.
[8] R. Fagin, A.O. Mendelzon and J.D. Ullman, A simplied universal relation assumption and its properties, *ACM Transactions on Database Systems* **7** (September 1982), 343–360.
[9] L. Geng and H.J. Hamilton, Interestingness measures for data mining: A survey, *ACM Computing Surveys* **38**(3) (2006), 9.
[10] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers Inc., 2005.
[11] A. Jiménez, F.Berzal and J.-C. Cubero, Frequent itemset mining in multirelational databases, in: *Proceedings of the 18th International Symposium on Foundations of Intelligent Systems*, 2009, pp. 15–24.
[12] A. Jimenez, F. Berzal and J.C. Cubero, POTMiner: Mining ordered, unordered, and partially-ordered trees, *Knowlegde and Information System* **23**(2) (2010), 199–224.
[13] J. Kleinberg and E. Tardos, *Algorithm Design*, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.
[14] F. Liu, Z. Lu and S. Lu, Mining association rules using clustering, *Intelligent Data Analysis* **5** (September 2001), 309–326.
[15] D. Maier and J.D. Ullman, Maximal objects and the semantics of universal relation databases,s *ACM Transactions on Database Systems* **8** (March 1983), 1–14.
[16] D. Maier, J.D. Ullman and M.Y. Vardi, On the foundations of the universal relation model, *ACM Transactions on Database Systems* **9** (June 1984), 283–308.
[17] J. Paterson, S. Edlich, H. Hörning and R. Hörning, *The Definitive Guide to db4o*, Apress, 2006.
[18] M. Plasse, N. Niang, G. Saporta, A. Villeminot and L. Leblond, Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set, *Computational Statistics and Data Analysis* **52**(1) (2007), 596–613.
[19] E.H. Shortliffe and B.G. Buchanan, A model of inexact reasoning in medicine, *Mathematical Biosciences* **23** (1975), 351–379.
[20] P.-N. Tan, V. Kumar and J. Srivastava, Selecting the right interestingness measure for association patterns, in: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, 2002, pp. 32–41.
[21] P.-N. Tan, V. Kumar and J. Srivastava, Selecting the right objective measure for association analysis, *Information Systems* **29** (June 2004), 293–313.
[22] G.I. Webb, Discovering significant patterns. *Machine Learning* **68** (July 2007), 1–33.
[23] X. Wu, D. Barbará and Y. Ye, Screening and interpreting multi-item associations based on log-linear modeling, in: *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, 2003, pp. 276–285.
[24] X. Yin, J. Han, J. Yang and P.S. Yu, CrossMine: efficient classification across multiple database relations, in: *Proceedings of the 20th International Conference on Data Engineering*, 2004, pp. 399–410.