Interestingness Measures for Association Rules within Groups

Aída Jiménez, Fernando Berzal, and Juan-Carlos Cubero

Dept. Computer Science and Artificial Intelligence,
ETSIIT - University of Granada, 18071, Granada, Spain
{aidajm,jc.cubero,fberzal}@decsai.ugr.es

Abstract. The study of association rules within groups of individuals in a database is interesting to define their characteristics and their behavior. In this paper, we define group association rules and we study interestingness measures for them. These evaluation measures can be used to rank groups of individuals and also rules within each group.

1 Introduction

Association rules have been used to analyze the relationships among the frequent itemsets in transactional and relational databases. Let $I = \{I_1, I_2, ..., I_m\}$ 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 $S \subseteq T$ [1]. An **association rule** is an implication of the form $A \Rightarrow C$, where $A \subseteq I$, $C \subseteq$ I , and $A \cap C = \emptyset$.

Databases can naturally contain groups of individuals that share some characteristics [2]. For example, in a census database, we can define groups of individuals according to their sex, their marital status, whether they have children, or even by combining several of such features. *Men who have children* are an example of such a group.

In this paper, we will describe how association rules can be defined in those groups to study the features that individuals in the same group have in common. For example, in the group of *men who have children*, we could find an association rule *lives in suburbs* $\Rightarrow$ *owns 2 car*, which is interpreted as *men who have children* and live in suburbs own 2 cars with some confidence value. If we obtained more rules, we could characterize the *men who have children* group and we could compare their behavior to the behavior of other groups in the database, for example, *men who are single.*

We define a **group** as a set of items $G = \{G_1, G_2, ..., G_n\}$ 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 this paper, we will describe how to adapt some of the interestingness measures that have been defined for association rules [3] [4] to group association rules and how these modified measures will help us to rank the different groups in a database in order to highlight the most interesting ones [5].

E. Hüllermeier, R. Kruse, and F. Hoffmann (Eds.): IPMU 2010, Part I, CCIS 80, pp. 298–307, 2010.

Our paper is organized as follows. In Section 2, we describe some rule evaluation metrics proposed in the literature. Section 3 introduces interestingness measures for group association rules. In Section 4, we explain how to order groups and group association rules within each group. Finally, we end our paper with some conclusions in Section 5.

2 Interestingness Measures for Standard Association Rules

The classical measures used to characterize an association rule are its support and its confidence [6][1].

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** s, where s is the percentage of transactions in D that contain $A \cup C$, i.e.,*

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

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

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

Confidence has some drawbacks as we can see in the example shown in Figure 1 where we have a graphical representation of two rules, $A \Rightarrow B$ and $A \Rightarrow C$. In the case of 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 for the $A \Rightarrow B$ rule is 75%. In the case of 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 also 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, however, the presence of the A reduces the presence of C, from 85% to 75%. Therefore, the confidence measure does not let us distinguish between these two cases.

In conclusion, confidence 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 alternatives to the support and confidence measures [3]. In the following paragraphs, we describe some of them:

Definition 4. *The interest of the rule $A \Rightarrow C$, also known as lift[7], is defined as:*

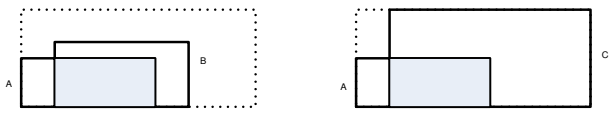

Fig. 1. Graphical depiction of two rules, $A \Rightarrow B$ and $A \Rightarrow C$, both with the same confidence but different consequent support

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

Interest measures 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 interest of the $A \Rightarrow B$ and $A \Rightarrow C$ rules in the the aforementioned example is $int(A \Rightarrow B)$=4.2 and $int(A \Rightarrow C)$=0.91. Here, $int(A \Rightarrow B) > int(A \Rightarrow C)$, which correspond to our intuition that $A \Rightarrow B$ is more interesting than $A \Rightarrow C$.

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

Definition 5. *The conviction[8] of the rule $A \Rightarrow C$ is defined as:*

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

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 interest measure.

In the example of Figure 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. In 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.

Unlike interest, rules that hold 100%, like the *Vietnam veteran* $\Rightarrow$ *more than 5 years old* rule, have the highest possible conviction value of ∞, which is an useful property. If 5% of the people are Vietnam veterans and 90% are more than five years old, then the interest of the *Vietnam veteran* $\Rightarrow$ *more than 5 years old* rule is (0.05)/(0.05)*0.9 = 1.11, sightly above 1, which is the value that would indicate statistical independence [8].

The main drawback of the conviction measure is that is not bounded, i.e., its range is [0,∞]. Therefore, it is difficult to establish a conviction threshold.

Let us now define 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:*

$$gain(A \Rightarrow C) = conf(A \Rightarrow C) - supp(C).$$

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

The gain values for the rules in Figure 1 is $gain\ (A \Rightarrow B)$=0.75-0.38=0.37 and $gain\ (A \Rightarrow C)$=0.75-0.85=-0.10. Figure 2 graphically shows these values. The length of the arrows represents 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 A is present.

Definition 7. *The certainty factor[9] of a rule $A \Rightarrow C$ is defined as:*

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

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

The certainty factor is the gain value normalized into the [-1,1] interval.

The certainty factor is interpreted as a measure of the variation of the probability that C is in a transaction when we consider only those transactions where A is. More specifically, a positive CF measures the decrease of the probability that C is not in a transaction, given that A is.

In the example of Figure 1 a), 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 the following paragraphs, we will explain how to adapt the measures described in Section 2 to group association rules, as well as how these new measures can be useful to evaluate this kind of association rules.

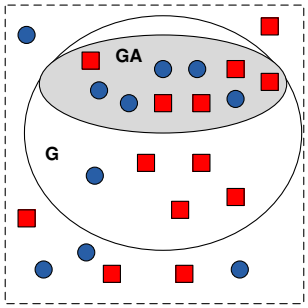

Fig. 3. Graphical representation of a group G

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.,*

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

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

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

$$supp_G(A \Rightarrow C) = \frac{P(GAC)}{P(G)} = conf(G \Rightarrow AC).$$

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

3.2 Group Confidence

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

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

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

3.3 Group Gain

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

$$\begin{aligned} gain_G(A \Rightarrow C) &= conf_G(A \Rightarrow C) - supp_G(C) \\ &= conf(GA \Rightarrow C) - conf(G \Rightarrow C) \end{aligned}$$

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 in the group.

In Figure 3, the support of the *circles* in the group G was $supp_G(\bullet) = 6/15 = 0.4$ 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 $gain_G(A \Rightarrow \bullet) = 0.5\text{-}0.4 = 0.1$. 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 to describe subgroups within the group G. For example, in the *men with children* group, the rule *lives in suburbs ⇒ 2 cars* might have a high positive gain value. That would suggest that, within the *men with children* group, living in suburbs increases the likelihood of owning

2 cars. Therefore, it would be easier to find men with two cars among those how live in suburbs than in the overall group of men who have children.

On the other side, rules with high negative gain values help us to find characteristics that do not define the subgroup. For example, in the *men older than 30* group, the rule *lives downtown* $\Rightarrow$ *owns a car* might have a negative gain. That would suggests that, within the *men older than 30* group, living downtown decrease the likelihood of owning a car. Therefore, it would be more difficult to find men older than 30 who own a car among those that live downtown than in the overall group.

Property 1. *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 Definition 11, 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 2. *The difference between the gain of the* $G : A \Rightarrow C$ *rule in the group* G *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 is $[-supp_G(C), 1 - supp_G(C)]$. In the followings paragraphs, we propose several ways to normalize the group gain depending on the kind of information we want to highlight. For example, 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.

Group Gain Factor

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

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

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

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

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 (C) is less frequent in G.

Group Variation

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 Figure 3 is $\delta_G(A \Rightarrow \bullet) = 0.1/(0.4) = 0.25$.

Group 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 of the group G that contain A (GA) given what we knew about G.

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 A in the group G.

In our example from Figure 3, $impact_G(A \Rightarrow \bullet) = (10)*0.1=1$. That should be interpreted as: there is 1 circle that we did not expect to be in GA when we only knew the support of $\bullet$ in G, $supp_G(\bullet)=0.4$, i.e., we did expect 4 circles in GA but there are 5, actually.

Impact Ratio

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 in 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 Figure 3.

4 Ranking Groups and Group Association Rules

The amount of rules and groups obtained in the rule mining process can be huge, hence it may be difficult to extract useful information from them. All these groups, as well as the rules within them, are obtained in an unsupervised process and we will use the measures we have described to highlight those rules and groups that might be relevant to the user according to several criteria.

In this section, we explain how to rank the groups and the rules within the groups according to their potential interestingness.

4.1 Ranking Rules within a Particular Group

The use of each measure provides us a different ordering among rules. We will choose a measure depending on the kind of information we want to highlight. In this section, we analyze how two rules in a group will have a different relative ordering in a group depending of the measure we use to evaluate them.

Characterizing subgroups within the 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.

- 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.
- 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.

Characterizing subgroups within the group using frequent features. If we are interested, not only in the subgroups, but also in using features that are frequent in our database, we should use a measure that takes into account the frequency of the antecedent of the rules, e.g., the **impact** measure.

This measure has the advantage that it has an easy interpretation: it indicates the number of individuals in G that are directly affected by the rule $A \Rightarrow C$, i.e., those individuals that are not expected to be in GA when we only know the overall support of C in the group.

4.2 Ranking Groups within the Database

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

include more interesting rules. For example, this can be useful for analyzing the behavior of groups in our database and studying the features that individuals in the same group have in common.

Impact seems to be a good measure to evaluate the interestingness of the group because it takes into account the number of individuals that are affected by each rule in the group. We should average the impact of the n rules 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. Then, they should not have the same weight.

We can define the weighted impact for the rules in a group using a different interest measure depending on the information we want to highlight. Formally, we define the weighted impact as:

$$\text{Weighted impact}(G) = \frac{\sum_{i=1}^{n} I_G(A \Rightarrow C) \cdot impact_G(A \Rightarrow C)}{\sum_{i=1}^{n} I_G(A \Rightarrow C)}$$

where $I_G(A \Rightarrow C)$ represents one of the interestingness measures described in Section 3 and analyzed in Section 4.1 for each $A \Rightarrow C$ rule in the group G.

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

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

5 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 are discovered within these groups of individuals.

We have adapted some of the standard interestingness measures for association rules to group association rules and we have also proposed new interestingness measures to evaluate this particular kind of association rules. We have studied the properties of these measures and which ones could be useful for the user depending on the information he is interested in.

Finally, we have proposed some guidelines to rank the groups in a database and the rules within each group depending on the user goals.

Acknowledgements

This work is supported by research project TIN2009-08296.

References

1. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco (2005)
2. Plasse, M., Niang, N., Saporta, G., Villeminot, A., Leblond, L.: Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set. Computational Statistics & Data Analysis 52(1), 596–613 (2007)
3. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3), 9 (2006)
4. Berzal, F., Blanco, I.J., Sánchez, D., Vila, M.A.: Measuring the accuracy and interest of association rules: A new framework. Intelligence Data Analysis 6(3), 221–235 (2002)
5. Bayardo Jr., R.J., Agrawal, R.: Mining the most interesting rules. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (1999)
6. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: 20th International Conference on Very Large Data Bases, pp. 487–499 (1994)
7. Brin, S., Motwani, R., Silverstein, C.: Beyond market baskets: generalizing association rules to correlations. In: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, vol. 26(2), pp. 265–276 (1997)
8. Brin, S., Motwani, R., Ullman, J.D., Tsur, S.: Dynamic itemset counting and implication rules for market basket data. In: Proceedings of the 1997 ACM SIGMOD international conference on Management of data, vol. 26, pp. 255–264 (1997)
9. Shortliffe, E.H., Buchanan, B.G.: A model of inexact reasoning in medicine. Mathematical biosciences 23, 351–379 (1975)