Weakness of Association Rules: A Mechanism for Clustering

Rajesh Natarajan, IT & Systems Group, Indian Institute of Management Lucknow, Lucknow - 226 013, Uttar Pradesh, India, T: +91-522-2736659, F: +91-522-2734025, rajeshn@iiml.ac.in

B. Shekar, Quantitative Methods & Information Systems Area, Indian Institute of Management Bangalore, Bangalore - 560 076, Karnataka, India, T: +91-80-26993093, F: +91-80-26584050, shek@iimb.ernet.in

ABSTRACT
We introduce the notion of weakness of an AR. After providing the intuition, we develop a weakness-based distance-function for clustering ARs. We cluster ARs obtained from a small artificial data set through the average-linkage method. The clusters are compared with those obtained by applying a commonly used method to the same data-set.

1. INTRODUCTION
Rule immensity is an important issue in Association Rule (AR) mining. This problem concerns the multitude of discovered rules that hinder easy comprehension. We define Weakness as the extent to which an AR is unable to explain the presence of its constituent items. Weakness is then used as a heuristic to group ARs. Rules with similar weakness are placed in the same cluster, thus facilitating easy exploration of connections among them. A user needs to examine only those rules in ‘relevant’ clusters.

2. WEAKNESS OF AN ASSOCIATION RULE
Consider an AR: \( R: a_1 \rightarrow \cdots \rightarrow a_n \rightarrow \cdots \rightarrow a_1 \) having support \( S_R \) and confidence \( C_R \). If all items of \( R \) are present in that transaction \( t \), then \( R \) covers \( t \). Let the support of an individual item \( a_i \) be \( R(t) \) with respect to database \( D \) be \( S_{a_i} \), \( R \) accounts for only \( \frac{S_{a_i}}{S_R} \) of the transactions in the database and does not explain the portion (of \( D \)) containing \( \frac{S_{a_i}}{S_R} \% \) of transactions containing \( a_i \). This fraction may be viewed as weakness of \( R \) with respect to its constituent \( a_i \):

\[
W_i = \frac{1}{S_{a_i}} \cdot \frac{S_{a_i}}{S_R} \cdot \{a_1, a_2, \ldots, a_n\}
\]

(1)

‘weakness’ brings out the strength of relationship between an AR and its constituents. A low \( w \)-value indicates significant characterization of its constituent items, since most of the transactions containing \( R \)’s constituent items exhibit the behaviour captured by \( R \). In addition, a low \( w \)-value signifies generality (wider coverage in \( D \)) of the relationship described by \( R \). In contrast, a high \( w \)-value indicates specificity of the relationships revealed by the rule.

3. A WEAKNESS-BASED DISTANCE MEASURE (\( d_w \))
Low generality of a high \( w \)-value rule suggests that relationships between the rule’s items and items present in other rules may be worth exploring. Actions taken only on the basis of a high \( w \)-value (high-specificity) rule could be skewed as the rule brings out only one aspect of the items’ behaviour in the database. Thus, weakness reflects the presence of relationships among constituents, action based on rules with equal or near-equal values could yield similar results.

We define weakness-based distance as:

\[
d_w(R_1, R_2) = \frac{\sum_{i=1}^{2} w_i \cdot w_j}{\sum_{i=1}^{2} w_i} \quad 0 \leq w_i, w_j \leq 1.
\]

(3)

Any difference \( \Delta w \) results in a larger distance for low \( w \)-values and smaller distance for high \( w \)-values. If \( \{w_1, w_2\} = \{w_1, w_j\} \) and \( w_i + w_j, w_i + w_j \), then \( d_w(R_1, R_2) \leq d_w(R_1, R_j) \). Let \( w = 0.4, w = 0.2, w = 0.8, w = 0.6 \). Then, \( d_w(R_1, R_2) = 0.3333 \) while \( d_w(R_1, R_4) = 0.14285 \). This may seem counter intuitive. However it has a rationale. \( R_1 \) are able to explain 40% and 20% respectively of their constituent items’ presence. Thus, they are more general than \( R_2 \) and \( R_4 \) whose \( w \)-values are 0.8 and 0.6 respectively. \( R_2 \) and \( R_4 \) have poorer explanatory power than \( R_1 \) and \( R_2 \) with respect to their constituent items.

This rationale has an analagous intuitive support. Consider four individuals \( A(R_1), B(R_2), C(R_3) \) and \( D(R_4) \). Assume \( A \) and \( B \) possess deeper knowledge (of a topic) than \( C \) and \( D \). Let the absolute difference in the knowledge-levels between the individuals in each of \( \{A,B\} \) and \( \{C,D\} \) be the same. Since \( A \) and \( B \) are quite knowledgeable, the difference would seem to be larger because it would require more effort to move from \( A \)’s knowledge-level to \( B \)’s knowledge-level. This greater effort may be due to the subtle and conceptually deeper knowledge required. However, it may be relatively easier to bridge the gap between \( C \) and \( D \). Fewer facts and straightforward knowledge acquisition may suffice. Similarly, \( R_1 \) and \( R_2 \) may have good explanatory power and hence they may be separated by a larger distance than the more specific pair \{\( R_3, R_4 \)\}.
Table 1. An artificial transaction dataset

<table>
<thead>
<tr>
<th>Transaction</th>
<th>Transaction</th>
<th>No.</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>{Bread,Butter}</td>
<td>{Bread,Jam}</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>{Bread,Milk}</td>
<td>{Bread,Jam}</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>{Milk,Chocolate}</td>
<td>{Biscuit,Biscuit}</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>{Milk,Chocolate,Biscuit}</td>
<td>{Butter,Milk}</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>{Pen,Pencil,Eraser}</td>
<td>{Pencil,Eraser}</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>{Milk,Biscuit,Pencil,Eraser}</td>
<td>{Bread,Butter,Milk}</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>{Bread,Jam,Milk}</td>
<td>--</td>
<td>12</td>
<td>--</td>
</tr>
</tbody>
</table>

Table 2. Association Rules extracted from transaction set of Table 1

<table>
<thead>
<tr>
<th>No</th>
<th>Rule</th>
<th>Support</th>
<th>Confidence</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>Butter→Bread</td>
<td>0.20</td>
<td>0.86957</td>
<td>0.321315</td>
</tr>
<tr>
<td>R2</td>
<td>Jam→Bread</td>
<td>0.21</td>
<td>1.00</td>
<td>0.243902</td>
</tr>
<tr>
<td>R3</td>
<td>Bread→Milk</td>
<td>0.30</td>
<td>0.7317</td>
<td>0.334146</td>
</tr>
<tr>
<td>R4</td>
<td>Butter→Milk</td>
<td>0.17</td>
<td>0.73913</td>
<td>0.460415</td>
</tr>
<tr>
<td>R5</td>
<td>Milk→Bread</td>
<td>0.14</td>
<td>0.82553</td>
<td>0.589947</td>
</tr>
<tr>
<td>R6</td>
<td>Chocolate→Biscuit</td>
<td>0.24</td>
<td>0.72727</td>
<td>0.136364</td>
</tr>
<tr>
<td>R7</td>
<td>Milk→Biscuit→Chocolate</td>
<td>0.11</td>
<td>1.00</td>
<td>0.662778</td>
</tr>
<tr>
<td>R8</td>
<td>Pen→Pencil,Eraser</td>
<td>0.13</td>
<td>0.8125</td>
<td>0.407738</td>
</tr>
<tr>
<td>R9</td>
<td>Pen→Pencil</td>
<td>0.13</td>
<td>0.8125</td>
<td>0.361607</td>
</tr>
<tr>
<td>R10</td>
<td>Pencil→Eraser</td>
<td>0.23</td>
<td>0.82143</td>
<td>0.146978</td>
</tr>
<tr>
<td>R11</td>
<td>Pencil→Eraser</td>
<td>0.16</td>
<td>1.00</td>
<td>0.192308</td>
</tr>
<tr>
<td>R12</td>
<td>Jam,Milk→Bread</td>
<td>0.16</td>
<td>1.00</td>
<td>0.509284</td>
</tr>
<tr>
<td>R13</td>
<td>Jam→Milk</td>
<td>0.16</td>
<td>0.76190</td>
<td>0.459048</td>
</tr>
<tr>
<td>R14</td>
<td>Chocolate→Milk</td>
<td>0.17</td>
<td>0.51515</td>
<td>0.572424</td>
</tr>
</tbody>
</table>

Table 3. \( d_{SC} \)-based clustering

<table>
<thead>
<tr>
<th>Step_No</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{R1,R2}</td>
</tr>
<tr>
<td>2</td>
<td>{R3,R4,R5}</td>
</tr>
<tr>
<td>3</td>
<td>{R6,R7,R8}</td>
</tr>
<tr>
<td>4</td>
<td>{R9,R10,R11,R12}</td>
</tr>
<tr>
<td>5</td>
<td>{R13,R14}</td>
</tr>
<tr>
<td>6</td>
<td>{R15}</td>
</tr>
<tr>
<td>7</td>
<td>{R16,R17}</td>
</tr>
<tr>
<td>8</td>
<td>{R18,R19,R20}</td>
</tr>
<tr>
<td>9</td>
<td>{R21,R22}</td>
</tr>
<tr>
<td>10</td>
<td>{R23,R24,R25,R26}</td>
</tr>
<tr>
<td>11</td>
<td>{R27,R28,R29}</td>
</tr>
<tr>
<td>12</td>
<td>{R30,R31,R32,R33}</td>
</tr>
<tr>
<td>13</td>
<td>{R34,R35,R36,R37,R38,R39,R40,R41}</td>
</tr>
</tbody>
</table>

Note: Values in the brackets represent merging distance.

It is easy to establish the metric properties of \( d_{SC}(R,R') \). The intuitive justification of \( d(R,R') \) and its being a metric enable \( d_{SC} \)-based clustering of ARs.

4. \( d_{SC} \)-BASED CLUSTERING OF ARs

Table 1 represents an artificial transaction database consisting of 100 transactions; the complete item-set being \{Bread,Butter,Jam,Milk,Chocolate,Biscuit,Eraser\}. It contains fifteen unique market-baskets. Support and confidence having respective thresholds of 0.1 and 0.5 yielded fourteen ARs listed in Table 2.

\( \text{R}_1 \) and \( \text{R}_2 \) have two common items namely, Chocolate and Biscuit. \( \text{R}_3 \) has a higher \( \text{v} \)-value. Support of \( \text{R}_4 \) (0.11) is much lower than that of \( \text{R}_5 \) (0.24). Hence \( \text{R}_5 \) is not able to account for the presence of Chocolate,Biscuit as much as \( \text{R}_4 \). Secondly, presence of Milk in \( \text{R}_5 \) further increases its \( \text{v} \)-value because \( \text{R}_5 \) is able to explain the presence of Milk in only 11 of the 50 transactions (22.0%) that contain Milk. However, a high support value does not guarantee a low \( \text{v} \)-value. \( \text{R}_5 \)'s \( \text{v} \)-value (Support=0.30,\( v=0.334146 \)) demonstrates this. \( \text{R}_5 \)'s support though high is not sufficient to cover the presence of Bread and Milk.

Table 3 lists the clusters obtained through the average-linkage method [4]. Despite the difference (0.017533) in the \( \text{v} \)-values between \( \text{R}_5 \) and \( \text{R}_8 \), being greater than the difference (0.010614) between \( \text{R}_5 \) and \( \text{R}_9 \), the former pair merges earlier. \( \text{R}_5 \) and \( \text{R}_8 \) being weaker rules leads to lesser inter-rule distance as compared to \( \text{R}_5 \) and \( \text{R}_9 \). A rule and its sub-rules being in different clusters may be due to the difference in support between a rule and its sub-rules. If the support values of a rule’s items have wide variation, then different sub-rules may explain their constituents’ presence to different extents. This difference in their \( \text{v} \)-values may place them in different clusters.

Cluster configuration after Step 9 results in clusters \( C_{w1}(R_2,R_5) \) and \( C_{w2}(R_5,R_7) \) whose elements have an average \( \text{v} \)-values of 0.608383 and 0.141671 respectively. \( \text{R}_5 \) is a member of high-\( \text{v} \)-value \( C_{w1} \) while its sub-rules \( R_8 \) and \( R_9 \) are members of clusters \( C_{w2} \) and low-\( \text{v} \)-value \( C_{w3} \) respectively. Support values of constituents Milk (0.50), Chocolate (0.33) and Biscuit (0.24) also show some variation. Thus, low-support coupled with high variation in the support values of its constituents might result in a weak rule.

Surprisingly, rules describing Milk (the most frequent item) belong to high-\( \text{v} \)-value clusters. None of the rules that contain Milk covers its presence to a substantial extent. High support of Milk also increases the \( \text{v} \)-value of low-support rules that contain it. Thus, a frequently occurring item may be present in many high-\( \text{v} \)-value rules if the item is purchased in many non-overlapping low-support market-baskets.

Another observation is with respect to clusters in clusters with relatively lower average \( \text{v} \)-values. Low-\( \text{v} \)-value clusters may not contain high-support rules. Consider \( C_{w3}(R_2,R_5) \). Note that support of \( R_5 \) (0.23) is quite close to support of its items \( \text{Pencil} \) (0.28) and \( \text{Eraser} \) (0.26). High explanatory power of such a rule is derived from its support value being close to the support values of its constituent items.

5. COMPARATIVE ANALYSIS AND DISCUSSION

Sahar [7] defines \( d_{SC} \)-distance on the basis of difference in rule’s itemsets and overlap in the set of transactions that each rule covers. \( d_{SC} \) considers itemsets in antecedent/consequent in their entirety while \( d \) considers each item of a rule separately. Table 4 displays \( d_{SC} \)-based cluster configurations.

\( R_5 \) is a sub-rule of \( R_8 \) both having support 0.13. Their antecedents match completely. Hence contribution due to antecedent dissimilarity towards \( d_{SC}(R_5,R_8) \) is 0. Also, \( R_5 \)’s consequent (\{Pencil\}) is a subset \( R_8 \)’s consequent (\{Pencil,Eraser\}). \( R_5 \) covers all transactions covered by \( R_8 \) thus increas-

Note: Values in the brackets represent merging distance.

Table 4. \( d_{SC} \)-based clustering

<table>
<thead>
<tr>
<th>Step_No</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>{R1,R2}</td>
</tr>
<tr>
<td>2</td>
<td>{R3,R4,R5}</td>
</tr>
<tr>
<td>3</td>
<td>{R6,R7}</td>
</tr>
<tr>
<td>4</td>
<td>{R8,R9}</td>
</tr>
<tr>
<td>5</td>
<td>{R10,R11,R12}</td>
</tr>
<tr>
<td>6</td>
<td>{R13,R14,R15,R16}</td>
</tr>
<tr>
<td>7</td>
<td>{R17,R18,R19,R20}</td>
</tr>
<tr>
<td>8</td>
<td>{R21,R22,R23,R24}</td>
</tr>
<tr>
<td>9</td>
<td>{R25,R26,R27,R28}</td>
</tr>
<tr>
<td>10</td>
<td>{R29,R30,R31,R32}</td>
</tr>
<tr>
<td>11</td>
<td>{R33,R34,R35,R36}</td>
</tr>
<tr>
<td>12</td>
<td>{R37,R38,R39,R40}</td>
</tr>
<tr>
<td>13</td>
<td>{R41,R42,R43,R44,R45,R46,R47,R48}</td>
</tr>
</tbody>
</table>

Note: Values in the brackets represent merging distance.
ing their similarity. Hence their low $d_{sc}$-value (0.429167). Hence $R_8$ and $R_9$ merge at Step 1.

$\Delta d_{sc}$-based clustering is useful in bringing together rules originating from the same portion of a database [7]. Here each cluster consists of rules whose items are members of the same or close domains. However, a rule and its sub-rules may vary a great deal on parameters like explanatory power, etc. Hence, a user may have to examine different clusters to find rules having the same specificity/generality.

Our scheme namely, groups rules having ‘similar’ values of $\text{weakness}$ (similar explanatory power) irrespective of their domain. Characteristics like average-$\text{weakness}$ may be used to define low-$\text{weakness}$ clusters leading to appropriate clusters for further examination. Rules in other clusters need not be examined thus mitigating the rule immensity problem to some extent.

6. REFERENCES


