Rule Power Factor: A New Interest Measure in Associative Classification

Ochina*, Suresh Kumarb, Nisheeth Joshic

*a,bManav Rachna International University, Sec 43, Faridabad-121001, India
b,cBanasthali Vidyapith, Nawai, Rajasthan – 304022, India

Abstract

In data mining it is generally anticipated that revealed knowledge should have characteristics of accuracy, reliability and interestingness. Most of the data mining algorithms find patterns that are accurate and reliable but might not be interesting. Interest measures are used to find the valued interesting rules which are useful to the user in effective decision making even in exceptional set of circumstances. A range of interest measures for rule mining have been suggested by researchers in the field of data mining to have different visualisations and analytics. In this paper, we have investigated a few of interest measures and proposed a new Interest measure with the name 'Rule Power Factor'. Experiments prove that this new interest measure is more informative and can act as a superset of 'Confidence' measure.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: Association rule mining, associative classification, lift, chi square, interest measure, data mining, rule power factor

1. Introduction

The field Associative Classification (AC) has tremendous opportunity and has many valued applications e.g. Market Basket Analysis, genetic epidemiological analysis, movies recommendation, heart disease prediction, cheminformatics, movies recommendation, ontology, bigdata, ubiquitous computing, Internet of Things. Associative Classification comprises with two techniques: Association rule mining (undefined) + Classification (Predefined).

* Corresponding author. Tel.: +91-0129-4198251.
E-mail address: ochinsharma2@gmail.com
Association rule mining is a very well-known technique by which various rules or association among variables can be defined. (Readers can refer paper [1] to read association rule mining in detail). Classification is a two phase process. In first phase, a sample data is collected from the data base and this is termed as training data. Using various constraints and techniques, a classification model is created to tag specific data with specific class label. When this model is justified on various parameters e.g completeness, accuracy, reliability on training data then test data is used to classify various data items with various predefined classes. Many researchers have proposed various interest measures for rule mining to fulfil the needs for effective decision making. Broadly, there are two types of classifications: objective and subjective. An objective measure has no awareness about the user. Mostly, objective measures are constructed on theories in probability, correlation and statistics. A subjective measure takes into account both the data and the user. This can be obtained by direct or indirect interaction with the user through the data mining process.

2. Related Work

There are various interest measures proposed by researchers. Each measure is basically to extract a specific pattern from the data. It depends on the need of pattern/data to be mined. For instance, user might be only interested in correlated data or to find the number of data items dominating in the database. At many occasions, finding negative rules are useful to avoid misleading decisions. For complete list of interest measures readers may refer paper [12]. Consider table 1, containing transactional data set used in sub-sections from 2.1 to 2.5 to understand basic terminology and a few well known interest measures.

<table>
<thead>
<tr>
<th>Table 1. Transactional data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>TID</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
</tbody>
</table>

2.1. Support

Support is defined as the probability that transactions in the database contains items both the antecedent and the consequent of the rule.

\[
\text{support (A→B)} = \frac{(\text{Transactions containing both A and B items})}{\text{Total number of transactions}}.
\]

\[
\text{support (A→B)} \Rightarrow \frac{2}{5} = 40\%
\]

If minimum threshold for support is chosen low, large numbers of rules are generated and evaluation of such rules is complex and time consuming. And choosing minimum threshold value high can make the pattern skipped and can compromise the effective decision making. Support, confidence, rule generation are explained in detail in [1].

2.2 Confidence

Confidence is a measure that finds out the association among antecedence and precedence part of a rule. It ignores the total number of transactions while calculating confidence.

\[
\text{confidence (A→B)} = \frac{\text{Total number of transactions containing items A and B}}{\text{Total number of transactions containing item A}}.
\]

\[
\text{confidence (A→B)} = \frac{2}{4} = 50\%
\]

2.3 Coverage

It measures how much database is covered by a rule A⇒ B. Hence, coverage is the number of transactions that satisfy the antecedent of a rule.

\[
\text{Coverage (A)} = \frac{\text{support (A)}}{\text{Total number of transactions}} = \frac{4}{5} = 0.800
\]
2.4. Leverage
Leverage measures the difference of actual and expected occurrences of A and B together in the data set and what is expected if A and B statistically dependent. The rational in a sales setting is to find out how many more occurrences are actual than expected from the independent sells [11].

\[
\text{Leverage (A->B)} = \text{confidence (A->B)} - \text{support (A) \cdot support (B)}
\]

\[
= \frac{2}{4} - \frac{3}{5} \times \frac{3}{5} = 0.020
\]

2.5. Lift/Interest
Lift was originally called interest. Lift measures how many times more often A and B occur together than expected if they were statistically independent. A lift value 1 indicates independence between X and Y [13].

Refer Table 1 data set to calculate lift.

\[
\text{lift (A->B)} = \frac{\text{support (A->B)}}{\text{support (A) \cdot support (B)}}
\]

\[
= \frac{\frac{2}{5}}{\left(\frac{2}{5}\right) \left(\frac{2}{5}\right)} = \frac{5}{6} = 0.83
\]

- If Lift (I) < 1, then A and B are said to be negatively interdependent.
- If I = 1, then A and B not found dependent and said to be independent of each other.
- If I > 1, then A and B appear more frequently together in the data and are said to be positively dependent on each other.

Limitations of lift / Interest:
For small values, the value of lift may vary unpredictably. In actual, the lift inclines to be higher for large itemsets as compare of small itemsets [2].

2.6. Chi-Square
The chi Square statistic is used to test correlation among various items. [4]. It is important to mention that confidence and correlation are two different things. Take an example

<table>
<thead>
<tr>
<th>Table 2. Transactional data set</th>
</tr>
</thead>
<tbody>
<tr>
<td>A     1</td>
</tr>
<tr>
<td>B     1</td>
</tr>
<tr>
<td>C     0</td>
</tr>
</tbody>
</table>

From this data one can depict that A and B are positively correlated, means most of the cases where A contains value 1, Y is also have same value 1 and vice a versa. Also A and C are negatively related. But after calculating support/confidence, we get A->B (support=25% and confidence 50%), A->C (support=37.5% and confidence 75%)

So, choosing a particular interest measure will change the importance of rules and hence impact on the decisions. Chi-square test has a condition that it should only be used when all cells in the contingency table have expected values greater than 1 and at least 80% of the cells have expected values greater than 5. The Chi-square test will produce larger values when the data set grows to infinity. The reason is that the Chi-square
value depends on the total number of transactions, whereas the critical cut off value only depends on the
degrees of freedom. It has been observed that many interest measures basically either not able to predict the
exact importance/strength of the rule or have different focus. So we are proposing a new interest measure that is
more focussed and abide with the rule importance/strength.

3. Proposed Interest Measure

The proposed algorithm for new interest measure named Rule Power Factor (RPF) is given here:

1. Generate frequent item sets with the desired minimal support. Observation: In a set, if \{A, B\} is a
frequent item set, then both A and B are frequent item sets too.
2. Find all n-item sets. Example (n=2): \(L_2 = \{ \{A,B\}, \{A,D\}, \{C,D\}, \{B,D\} \}\)
3. Generate (n+1)-item sets by merging n-item sets. \(L_3 = \{ \{A,B,C\}, \{A,C,D\},\n\{A, B, D\}, \{B, C, D\} \}\).
4. Test the newly generated (n+1)-items set for minimum support.
5. Eliminate sets/rules <minimum support
6. Test the remaining item sets for minimal support by counting their
occurrences in data.
7. Increment \(n\) and continue until no more frequent item sets can be generated.
8. Check all the generated sets/rules against minimum confidence.
   \(\text{confidence}(A\rightarrow B) = \text{support}(A,B) / \text{support}(A)\)
9. Discard rules < minimum confidence
10. Next, compute the power factor for all the shortlisted rules.
11. Rule Power Factor (RPF) = Rule confidence* Rule support => \(\text{confidence}(A\rightarrow B) * \text{support}(A\rightarrow B)\)

3.2. RPF significance over confidence measure

RPF focuses on the importance of association between antecedent and consequent of rules. RPF works well
even where confidence fails:

(a) If item A appeared in 20 transactions and B in 50 out of total 100 transactions and item A and B both
together appear 15 transactions. Then \(\text{conf}(A\rightarrow B) = .15/.2 = 0.75 = 75\%\).

(b) If item A appeared in 30 transactions and B in 60 out of total 100 transactions and item A and B both
together appear 20 transactions. Then \(\text{conf}= .2/.3 = .66 = 66\%\).

But, in case (b), both antecedent and consequent item’s occurrences increased individually and with together.
While interest measure confidence says surprisingly that case (a) is more important than case (b). If we take the
help of Rule Power Factor (RPF):

RPF: confidence(A\rightarrow B)*sup(A)
(a) 0.75*0.15= 0.11
(b) 0.66*0.2= 0.13
RPF, correctly judge that case (b) is more important.

3.3. RPF significance over Lift measure

RPF works well even where Lift (a well-known and accepted interest measure) fails:

(a) If item A appeared in 20 transactions and B in 50 out of total 100 transactions and item A and B both
together appear 15 transactions.
Lift = \(0.15/(0.2*0.5)=0.15/0.10=1.15\)

When in the 100 transaction data base, item A and B together occur 15 times, lift says positive for the rule.

(b) If item A appeared in 30 transactions and B in 60 out of total 100 transactions and item A and B both together appear 20 transactions.
Lift=\(0.2/(0.3*0.6)=0.2/0.18=1.11\)

Surprisingly, when in the same 100 transaction data base, item A and B together occur 30 instead 15, times case b), lift says rule a is important.

Now let’s see RPF for both cases
RPF: \(\text{conf}(A \rightarrow B) \times \text{sup}(A)\), Case (a) RPF=0.75*0.15=0.11, Case (b) RPF=0.15*0.2 =0.30, 0.66*0.2=0.13

RPF rightly predicted that case (b) is more important than case (a)

3.4. Analysis of Piatetsky-Shapiro Principles for RPF

Piatetsky-Shapiro [11] proposed three principles for interest measure F

- P1. F = 0 if A and B are statistically independent; i.e., \(P(AB) = P(A) P(B)\).
  The first principle states that an association rule that occurs by chance has zero interest value of measure, i.e., it cannot be interesting. If, for a scenario, \(P(A)=20/100=.2, P(B)=10/100=.1, P(AB) = 2/100=.02\),
  \(\text{conf}(A \rightarrow B)=.02/.2=0.1\)
  \(P(AB) = P(A) P(B)\).
  \(0.02=.2*.1=0.02\) Now calculate F i.e RPF, \(\text{conf}(A \rightarrow B) \times \text{sup}(A \rightarrow B) =0.1*.02=.00=0\)

- P2. F monotonically increases with \(P(AB)\) when \(P(A)\) and \(P(B)\) remain the same.
  \(P (A) = 20 / 100 = .2, P (B) = 10 / 100 = .1, P (AB) = 5 / 100 = .05, \text{conf} = P (AB) / P (A) = .05 / .2 = .25,\)
  \(\text{RPF} = .25 * .05 = .0125,\)
  Now, if \(P (AB) = 10 / 100 = .1, P (B) = 10 / 100 = .1, P (A) = 20 / 100=.2, \text{conf} = P (AB) / P (A) = .1 / .2 = .5, \text{RPF} = .5 * .1 = .05\) Hence, RPF obeys principle P2.

- P3. F monotonically decreases with \(P(A)\) (or \(P(B)\)) when \(P(AB)\) and \(P(B)\) (or \(P(A)\)) remain the same[11].
  The third principle states that if the supports for \(A \rightarrow B\) and \(B \rightarrow A\) are fixed, the smaller the support for \(A\) (or \(B\)) is the more interesting the pattern is.[2]
  \(P (A) = 20 / 100 = .2, P (B) = 10 / 100 = .1, P (AB) = 5 / 100 = .05, \text{conf} = P (AB) / P (A) = .05 / .2 = .25,\)
  \(\text{RPF} = .25 * .05 = .0125,\)
  if \(P (AB) = 5 / 100 = .05, P (B) = 10 / 100 = .1, P (A) = 10 / 100 = .1, \text{conf} = P (AB) / P (A) = .05 / .1 = 0.5,\)
  \(\text{RPF} = 0.5 * .05 = .025,\)
  Hence, RPF follow principle P3.

3.5. Analysis of Tan’s Principles for RPF

Tan et al. proposed five properties based on operations on 2x2 contingency tables [12].

- O1: F is symmetric under variable permutation.
- O2: F is the same when we scale any row or column by a positive factor.
- O3: F becomes –F under row/column permutation.
- O4: F remains the same under both row and column permutation.
- O5: F has no relationship with the count of the records that do not contain A and B.
RPF is an asymmetric Interest Measure. Asymmetric measures have their own logical importance.

Let’s take an example, if an item A→B has confidence 90% and item B→A = 40% i.e anyone who purchases item A 90% purchases item B, but those purchase item B, purchases only item A are 40%. So in sales promotions when customers buy item B, should be given discount on purchase of item A but vice a versa is not needed.

In [7], Hilderman et al. suggested five principles as a good measure should follow. Some of these principles are alike to the ones offered by Piatetsky-Shapiro, whereas others may not be appropriate to association analysis as they take up that the count attribute values are in sorted order (such assembling is less intuitive for contingency tables). That’s why no rule interest measure could obey all the 5 principles.

Table 3. Analysis of Piatetsky-Shapiro and Tan’s Principles for various interest measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>P1</th>
<th>P2</th>
<th>P3</th>
<th>O1</th>
<th>O2</th>
<th>O3</th>
<th>O4</th>
<th>O5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Confidence</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Lift/Interest</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Coverage</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>Leverage</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>RPF</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y*</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

4. Experiment and Observation

Experiment is conducted using Weka tool 3.7.13 version with windows 7 operating system. Intel core i3, 2.40GHz processor, 4GB Ram. Weka is an open source tool. We have used JEdit to edit Weka’s code written in Java and ANT software to recompile the Weka source JAR file after coding. We used 10 best rules after simulating data with default parameters setting on weather dataset. RPF calculation and Result Analysis is shown below in ‘Associator output’ window:

![Fig 1: Weka Experiment](image-url)
As we can see RPF is more informative regarding importance of rules. When antecedent and consequent association increases, rule importance increases and hence RPF, whereas other built in measures in Weka couldn’t achieve it. However for the rules having same values of RPF; rule ranking techniques can be applied e.g random pick, sequential pick, based on timestamp, based on the number of attributes in antecedent and consequent.

5. Conclusion

In this paper, we have developed a new interest measure to judge the importance or strength of a rule. Our interest measure works well even where support, confidence, lift, chi square and other measures fails. Our interest measure has also qualified the good interest measure principles postulated by Piatetsky Shapiro. So, using RPF, data mining area can be benefited by driving the important rules for decision making and avoid misleading rules and decisions.

Reference