Interestingnesslab: A Framework for Developing and Using Objective Interestingness Measures

Lan Phuong Phan[1(✉)], Nghia Quoc Phan[2], Ky Minh Nguyen[3], Hung Huu Huynh[4], Hiep Xuan Huynh[1], and Fabrice Guillet[5]

[1] Can Tho University, 3/2 Street, Ninh Kieu District, Can Tho City, Vietnam
{pplan,hxhiep}@ctu.edu.vn
[2] Tra Vinh University, No. 126 National Road 53, Ward 5, Tra Vinh City, Vietnam
nghiatvnt@gmail.com
[3] Can Tho University of Technology,
256 Nguyen Van Cu Street, Ninh Kieu District, Can Tho City, Vietnam
nmky@ctuet.edu.vn
[4] University of Science and Technology - University of Danang,
54 Nguyen Luong Bang Street, Lien Chieu District, Da Nang City, Vietnam
hhhung@dut.dun.vn
[5] Polytech Nantes, University of Nantes,
La Chantrerie rue Christian Pauc BP 50609, 44306 Nantes Cedex 3, France

Abstract. The objective interestingness measures play an important role in data mining because they are used for mining, filtering and ranking the patterns. However, there is no research that collects the measures fully as well as there is no tool that can: automatically calculate the interestingness values of the patterns by using those measures, and is the framework for rapidly developing the applications related to objective interestingness measures. This paper describes Interestingnesslab - a tool of the objective interestingness measures is developed in the R language. The main functions of the tool are: mining a set of association rules and presenting them by the cardinalities $(n, n_X, n_Y, n_{X\overline{Y}})$, calculating the interestingness value of an association rule according to 1 of 109 collected measures; calculating the interestingness values of the whole rule set in many measures selected by the user; discovering the tendencies in a data set and recommending the top N items to the user; and studying the specific behavior of a set of interestingness measures in the context of a specific dataset and in an exploratory data analysis perspective. With Interestingnesslab, the user can easily and quickly reuse its functions to develop his/her own applications.

Keywords: Objective interestingness measure · Interestingnesslab · Association rule · Recommender system

1 Introduction

The data mining process takes a data set as the input and generates the patterns (such as the association rules, the classification rules) as the output [5].

M. Akagi et al. (eds.), *Advances in Information and Communication Technology*,
Advances in Intelligent Systems and Computing 538, DOI 10.1007/978-3-319-49073-1_33

In fact, the data mining process can create hundreds and thousands of patterns. The determination of the most useful patterns can be performed by using the interestingness measures to calculate the actual value of the patterns. The interestingness measures play an important role in mining data, regardless of the type of patterns. They can be used for: (1) - pruning the unattractive patterns during the data mining process to narrow the search space and thus improve the efficiency of mining. For example, a threshold on the support measure can be used to remove patterns with low support values during mining process; (2) - ranking the patterns according to their interestingness values; (3) - filtering the interesting patterns during the post-processing. If the interestingness measures are good, the cost of time and space in mining data will be reduced. Each interestingness measure characterizes a certain aspect of the data set, therefore, the users should select the appropriate measure meeting their needs, calculate the interesting values of the patterns in the selected measure, and then extract the useful patterns.

The interestingness measures can be divided into two categories: subjective measures and objective measures [2,11]. The subjective approach evaluates the patterns by using the target, the knowledge, and the belief of user. The objective approach uses the statistical characteristics of the patterns to evaluate the interestingness. The second approach is only based on the raw data and does not require knowledge on the users or the application. Most interestingness measures are the objective interestingness measures. The objective interestingness measures are studied, surveyed by many independent group of authors, and at different times, such as Tan et al. in 2004 [10], Geng et al. in 2006 [1], Huynh et al. in 2008 [1], Heravi et al. in 2010 [8]; Grissa et al. in 2012 [7]; and Tew et al. in 2014 [12]. However, these studies just focus on the measures suitable for their own research orientation, and often focus on the common measures. For example, Huynh at el ranked 40 objective interestingness measures with sensitivity values; and Tew et al. focused on an analysis of the rule-ranking behavior of 61 well-known interestingness measures.

Although, there are a lot of researches on the interestingness measures, there still exist some mistakes in some researches: (1) - cite the formula of some measures incorrectly (the formula is improper as it is presented in the original research); (2) - use a measure that is called by different names, but just mention one name and do not take a note (or do not know) the remaining names. The mistake or the omission could be repeated if the latter researches refer to and cite from the previous researches, thereby affecting the quality of research. Besides, at the present, there is no research that synthesizes the objective interestingness measures fully, especially the recently proposed measures. The synthesize of the objective interestingness measures will form a common, complete, and reliable reference system which enables the researchers to save a lot of time and effort when studying the association rules and the measures of data mining. Moreover, there is also no automatic tool that meets the following criteria: (1) - calculate the value of each association rule according to many objective interestingness measures; (2) - is created as a framework for quickly developing applications to

detect the useful patterns, and then these applications can be easily integrated to the tool; (3) - is developed in R, a language and environment for the statistical computing and graphics. From this analysis, we propose a tool, named Interestingnesslab, to aggregate objective interestingness measures fully as well as provide the main functions as the framework for developing and using the objective interestingness measures.

This paper is organized into 5 parts. The first part is the introduction. The second part presents interestingness values. The third part describes the overview architecture of Interestingnesslab. The fourth part is core functions of the tool. The last part concludes this paper.

2 Interestingness Values

2.1 Objective Interestingness Measures

The objective interestingness measures used for evaluating the quality of patterns (i.e. the association rules in this paper) use statistics derived from data to determine whether an association rule is interesting. As mentioned in Part 1, there is no research that synthesizes the objective interestingness measures fully.

To collect the objective interestingness measures effectively, some criteria are identified: (1) - be the objects of researches on the interestingness measures as well as be cited by many others papers, (2) - be published by the reliable sources such as IEEE, Springer, ACM, Science Direct; (3) - be independently studied by the groups of authors.

After being collected, analyzed and validated, there are 109 different objective interestingness measures (109 different formulae), and 21 groups in which each group consists of some measures called by different names but having the same formula (Appendix). Formulae will be used for calculating the interestingness value of the association rules.

2.2 Presentation of an Association Rule

Let $I = \{I_1, I_2, \ldots, I_m\}$ be the set of different attributes (items); $D = \{T_1, T_2, \ldots, T_n\}$ be a transaction database in which each record $T_i(i : 1 \ldots n)$ is a transaction, and T_i is a subset of items ($T_i \subseteq I$), an association rule [11] is denoted by $X \rightarrow Y$ where X is called antecedence, Y is called consequence, X and Y are the subsets of items, and $X \cap Y = \emptyset$. An association rule represents the implicative trend between the item sets.

The presentation of an association rule $X \rightarrow Y$ can be expressed by a set of 4 values n, n_X, n_Y, and $n_{X\overline{Y}}$. $\{n, n_X, n_Y, n_{X\overline{Y}}\}$ is called the cardinality of an association rule where n is the number of transactions; $n_X = card(X)(n_Y)$ is the number of transactions that have $X(Y)$; and the counter-example number $n_{X\overline{Y}} = card(X \cap \overline{Y})(\overline{Y}$ is the complementary set of $Y)$ is the number of transactions that have X but do not have Y(Fig. 2).

For example, the association rule $\{egg, meat\} \rightarrow \{beer\}$ mined from the data set in Fig. 1 is represented by the cardinality $\{5, 3, 3, 1\}$.

	Bread	Egg	Meat	Beer	Milk
T_1	1	1	0	0	0
T_2	1	0	1	1	1
T_3	0	1	1	1	0
T_4	1	1	1	1	0
T_5	1	1	1	0	0

Fig. 1. An example of a transaction database.

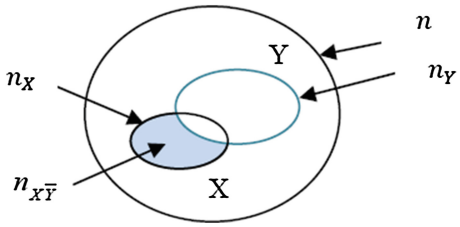

Fig. 2. The presentation of an association rule $X \rightarrow Y$.

2.3 Interestingness Value

The formula of an objective interestingness measure can be expressed by a function of 4 parameters n, n_X, n_Y, and $n_{X\overline{Y}}$: $m(X,Y) = f(n, n_X, n_Y, n_{X\overline{Y}})$. For example, the formula of the measure Support is $\frac{n_X - n_{X\overline{Y}}}{n}$. For 109 collected measures, their formulae are written in many different forms, such as the frequency, the number of transactions, etc. Therefore, for the convenient, all those formulae are converted to the functions of the cardinality n, n_X, n_Y, and $n_{X\overline{Y}}$.

The interestingness value (the quality) of an association rule $X \rightarrow Y$ in a measure is calculated by using the formula of that measure and the presentation of the rule $X \rightarrow Y$ (the cardinality $\{n, n_X, n_Y, n_{X\overline{Y}}\}$).

For example, if the association rule $\{egg, meat\} \rightarrow \{beer\}$ mined from the data set in Session 2.2 is represented by the cardinality $\{5, 3, 3, 1\}$, the interestingness value of this rule in the measure Support is $\frac{n_X - n_{X\overline{Y}}}{n} = \frac{3-1}{5} = 0.4$.

3 Architecture of Interestingnesslab

The overview architecture of Interestingnesslab is displayed as Fig. 3. The main components of this tool are: *cardinality*, *utility*, *application*, *interestingnessvalues*, and *interestingnessmeasures*.

The component *cardinality* is responsible for calculating the cardinalities of the rule set. It takes an association rule set generated by the Apriori algorithm, and a data set as the inputs; and generates the matrix *cardinality_matrix* as the output. Each row of *cardinality_matrix* includes the information: the ordinal

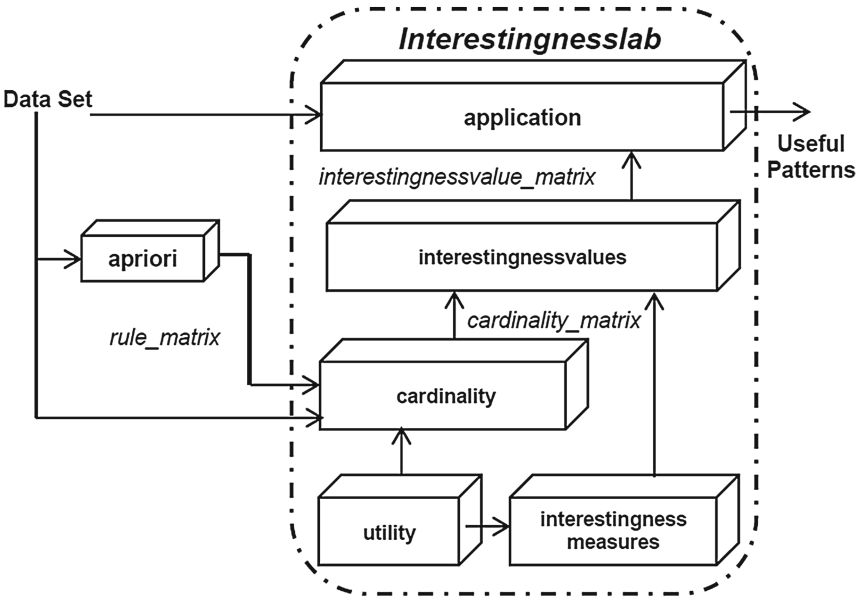

Fig. 3. The overview architecture of Interestingnesslab.

#	n	n_X	n_Y	$n_{X\overline{Y}}$	Presentation of rule
...	...	...	...	...	...
30	93	28	54	12	{Calculus} => { Probability and Statistics }
...	...	...	...	...	...

Fig. 4. An example of the matrix *cardinality_matrix*.

number (of a rule), $n, n_X, n_Y, n_{X\overline{Y}}$, the presentation of a rule in form $X \rightarrow Y$. Figure 4 shows an example of the matrix *cardinality_matrix*.

The component *utility* is a set of the utility functions that are used by the component *cardinality*.

The component *interestingnessvalues* is responsible for calculating the interestingness values of a rule set in the selected measures. This component takes *cardinality_matrix* as the input; generates *interestingnessvalue_matrix* as the output. Each row of *interestingnessvalue_matrix* consists of the information: the ordinal number (of a rule), $n, n_X, n_Y, n_{X\overline{Y}}$, the presentation of a rule in form $X \rightarrow Y$, the interestingness value of the first selected measure, the interestingness value of the second selected measure, etc. Figure 5 shows an example of the matrix *interestingnessvalue_matrix*.

The component *interestingnessmeasures* is a set of the functions where each function gets 4 parameters $n, n_X, n_Y, n_{X\overline{Y}}$ representing for an association rule; and returns the interestingness value of that association rule in a

#	n	n_X	n_Y	$n_{X\overline{Y}}$	Presentation of rule	Implication Intensity	Rule Interest
...	...	...	...	...	...	...	...
30	93	28	54	12	{Calculus} => { Probability and Statistics }	0.394548	-0.25806
...	...	...	...	...	...	...	...

Fig. 5. An example of the matrix *interestingnessvalue_matrix*.

specific measure. The function name is the measure name. The functions of *interestingnessmeasures* is used by the component *interestingnessvalues*.

The component *application* is an open component including the applications that are built by the users themselves as well as by four above components. At present, there are two applications already developed in this component: *ARQAT* and *ARbasedRS*. *ARQAT* (Association Rule Quality Analysis Tool) studies the specific behavior of a set of the interestingness measures in the context of a specific dataset and in an exploratory data analysis perspective. This tool implements 14 graphical and complementary views structured on 5 levels of analysis: ruleset analysis, correlation and clustering analysis, best rules analysis, sensitivity analysis, and comparative analysis. *ARQAT* was first developed in Java by Huynh et al. [9]. To integrate this tool to Interestingnesslab, *ARQAT* is re-implemented in R. The detail description of this tool is presented in [9]. Therefore, this paper does not remind the functions of *ARQAT*. *ARbasedRS* (Association Rule based Recommender System) discovers tendencies in a data set, and recommends the top N items to a user.y

4 Some Core Functions of Interestingnesslab

4.1 Presenting a Rule Set in the Form $\{n, n_X, n_Y, n_{X\overline{Y}}\}$

An association rule $X \rightarrow Y$ can be represented by a cardinality $\{n, n_X, n_Y, n_{X\overline{Y}}\}$. The following algorithm shows how to calculate $n, n_X, n_Y, n_{X\overline{Y}}$ for each rule of the rule set.

The algorithm for calculating the cardinalities of a rule set

Input: a set of the association rules (*ruleset*), a transaction database (*data*).
Output: the cardinalities of a set of rules (*cardinality_matrix*).
Steps:
- Count the number of transactions n;
- Calculate n_X

  Transform the left hand side of *ruleset* into a matrix *lhsRules* in which $lhsRules[i, j] = TRUE$ if the item j is one element of the left hand side of rule i, and $lhsRules[i, j] = FALSE$ otherwise;
  Calculate the matrix (cross) product $lhsProduct = lhsRules * t(data)$;
  Count n_X of each rule i $n_X[i] = rowSum(lhsProduct[i])$;
- Calculate n_Y. The method for calculating n_Y is similar to the method for

calculating n_X, except that it uses the right hand side of *ruleset*;
- Calculate $n_{X\overline{Y}}$:
    Calculate n_{XY}. The method for calculating n_{XY} is similar to the method for calculating n_X, except that it uses both side of *ruleset*;
    Calculate $n_{X\overline{Y}}$ of each rule i: $n_{X\overline{Y}}[i] = n_X[i] - n_{XY}[i]$;
- Concatenate n, n_X, n_Y, and $n_{X\overline{Y}}$ with *ruleset* to create the matrix *cardinality_matrix*.

4.2 Calculating the Interestingness Value of an Association Rule

Using 109 formulae of the objective interestingness measures converted to $\{n, n_X, n_Y, n_{X\overline{Y}}\}$, 109 functions are implemented. Each function takes the values n, n_X, n_Y, and $n_{X\overline{Y}}$ representing for an association rule as the input, and returns an interestingness value of that rule as the output.

4.3 Calculating the Interestingness Value of a Rule Set

Instead of calculating the interestingness value of an association rule in a measure, this function allow a user to calculate the interestingness values of a rule set in selected measures.

The algorithm for calculating the interestingness values of a rule set

Input: the cardinalities of a rule set (*cardinality_matrix*), a list of selected measures (*measures*).

Output: the interestingness values of a rule set in selected measures (*interestingnessvalue_matrix*).

Steps:

- Set ruleNum = the number of rules (rows of *cardinality_matrix*);
- Set measureNum = the number of the selected measures;
- Calculate the interestingness values:

```
for (i=1; i≤ruleNum; i++)
    Access the cardinality of rule i:
    for (j=1; j≤measureNum; j++)
        Calculate the interestingness value of rule i in measure
        j with name m: v = m(n, n_X, n_Y, n_{XȲ})
        Write the value v to the matrix value: values[i][j] = v
    }
}
```

- Concatenate two matrices (*values* and *cardinality_matrix*) to return the matrix *interestingnessvalue_matrix*.

4.4 Discovering Tendencies and Recommending Top N Items

The application called Association Rule based Recommender System is implemented by using the above functions. This system is developed to discover the tendencies in a data set, and recommend the top items to a user.

The algorithm for discovering the tendencies and recommending the top N items

Input: a transaction database (*data*) including a subset of items T_a that u_a liked.

Output: the tendencies or the top N items.

Steps:

- Generate a set of rules (*ruleset*) by using the Apriori algorithm. The user can set the thresholds on two measures support and confidence;
- Calculate the cardinalities $\{n, n_X, n_Y, n_{X\overline{Y}}\}$ of the rule set;
- Select the objective interestingness measures that are suitable for the user's purpose;
- Calculate the interestingness value $c(X, Y)$ for each rule of the *ruleset* in the selected measure;
- Sort the rules with their interestingness values in the descending order;
- Select 2 options:

  Show the tendencies in a data set by using the threshold on the selected measure.

  Recommend to a user u_a the top N items that u_a can like:

  Find all matching rules $X \rightarrow Y$ for which $(X \subseteq T_a)$;

  Recommend N right hand sides (Y) of matching rules with the highest values.

5 Conclusion

This paper has collected and validated 109 objective interestingness measures, then converted their formula to the unified format (the cardinality $\{n, n_X, n_Y, n_{X\overline{Y}}\}$). The list of these measures can be regarded as a complete, systematic, and reliable reference source. Besides, the tool of the objective interestingness measures, named Interestingnesslab, has been developed with the main functions: presenting an association rule set by the cardinalities; calculating the interestingness values of a rule in a specific measure; calculating the interestingness values of the rule set in measures selected by the user; building an application to detect the tendencies in a data set and to recommend the top N items to a user; and studying the specific behavior of a set of the interestingness measures in the context of a specific dataset and in an exploratory data analysis perspective. Interestingnesslab is implemented in the R language, and is an open source package. Therefore, the users can fully reuse the core functions to develop and use their own applications.

Appendix

21 groups of measures, each group includes the measures called by the different names but having the same formula.

#	Group of measures	#	Group of measures
1	Accuracy, Causal Support	2	Added Value, Pavillon, Centered Confidence
3	Bayes Factor, Odd Multiplier	4	Correlation Coefficient, Phi-Coefficient, Pearson's Correlation Coefficient, Linear-Correlation, Newrelevancy
5	Cosine, Ochia, IS Measure	6	Descriptive Confirmed-Confidence, Lerman Similarity Index
7	Dice Index, Czekanowski Dice, F-Measure Examples and Contra-Examples	8	Directed Contribution to Chi square, Lerman Similarity Index
9	Rate, Example and Contra-Example Rate, Encountered Rate	10	Gray and Orlowska's Interestingness Weighting Dependency, I-Measure
11	Indice Probabilistc d'Ecart d'Equilibre, Probabilistic Measure of Deviation from Equilibrium(IPEE)	12	Jaccard, Coherence
13	Kappa Coefficient, Cohen	14	Kulczynski 1, Agreement-Disagreement Index
15	Lift, Interest	16	Loevinger, Certainty Factor, Satisfaction
17	Mutual Information, 2-way Support Variation	18	Normalized Difference, Match
19	Piatetsky-Shapiro, Pearl, Leverage 2, Carnap, Novelty	20	Relative Risk, Class Correlation Ratio
21	Specificity 1, Negative Reliability		

References

1. Huynh, H.X., Guillet, F., Le, T.Q., Briand, H.: Ranking objective interestingness measures with sensitivity values. VNU J. Sci. Nat. Sci. Technol. **24**, 122–132 (2008)
2. McGarry, K.: A survey of interestingness measures for knowledge discovery. Knowl. Eng. Rev. J. **20**(1), 39–61 (2005)
3. Guillet, F., Hamilton, H.Z.: Quality Measures in Data Mining. Series in Computational Intelligence, vol. 43. Springer, Heidelberg (2007)
4. Tan, P.-N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Pearson (2006)
5. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: a survey. ACM Comput. Surv. **38**(3) (2006). Article 9
6. Guillaume, S., Grissa, D., Nguifo, E.M.: Categorization of interestingness measures for knowledge extraction. CoRR abs/1206.6741 (2012)
7. Grissa, D., Guillaume, S., Nguifo, E.M.: Combining Clustering techniques and FCA to characterize Interestingness Measures, Research Report LIMOS/RR-12-05 (2012)

8. Heravi, M.J., Zaïane, O.R.: A study on interestingness measures for associative classifiers. In: Proceedings of the 2010 ACM Symposium on Applied Computing, SAC 2010, pp. 1039–1046 (2010)
9. Huynh, X.H., Guillet, F., Briand, H.: ARQAT: an exploratory analysis tool for interestingness measures. In: Proceedings of the 11th International Symposium on Applied Stochastic Model and Data Analysis, ASMDA 2005, Brest, France, pp. 334–344 (2005)
10. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Trans. Knowl. Data Eng. **8**(6), 970–974 (1996)
11. Tan, P.-N., Kumar, V., Srivastava, J.: Selecting the right objective measure for association analysis. Inf. Syst. **29**(4), 293–313 (2004)
12. Tew, C., Giraud-Carrier, C., Tanner, K., Burton, S.: Behavior-based clustering and analysis of interestingness measures for association rule mining. Data Mining Knowl. Discov. **28**(4), 1004–1045 (2014). Springer, US