Expected vs. Unexpected: Selecting Right Measures of Interestingness

Chapter · September 2020

DOI: 10.1007/978-3-030-59065-9_4

5 authors, including:


Rahul Sharma
Tallinn University of Technology Estonia


Minakshi Kaushik


Sijo Pious
Tallinn University of Technology


Sadok Ben Ben Yahia
University of Tunis El Manar

Expected vs. Unexpected: Selecting Right Measures of Interestingness

Rahul Sharma[1](✉), Minakshi Kaushik[1], Sijo Arakkal Peious[1], Sadok Ben Yahia[2], and Dirk Draheim[1]

[1] Information Systems Group, Tallinn University of Technology, Akadeemia tee 15a, 12618 Tallinn, Estonia
{rahul.sharma,minakshi.kaushik,sijo.arakkal,dirk.draheim}@taltech.ee
[2] Software Science Department, Tallinn University of Technology, Akadeemia tee 15a, 12618 Tallinn, Estonia
sadok.ben@taltech.ee

Abstract. Measuring interestingness in between data items is one of the key steps in association rule mining. To assess interestingness, after the introduction of the classical measures (support, confidence and lift), over 40 different measures have been published in the literature. Out of the large variety of proposed measures, it is very difficult to select the appropriate measures in a concrete decision support scenario. In this paper, based on the diversity of measures proposed to date, we conduct a preliminary study to identify the most typical and useful roles of the measures of interestingness. The research on selecting useful measures of interestingness according to their roles will not only help to decide on optimal measures of interestingness, but can also be a key factor in proposing new measures of interestingness in association rule mining.

Keywords: Knowledge discovery in databases · Association rule mining · Measures of interestingness

1 Introduction

In knowledge discovery in data (KDD), association rule mining (ARM) is one of the most established data mining techniques. It is commonly used to find out interesting patterns between data items in large transactional data sets. In ARM, association rules are accompanied by measures of interestingness (support, confidence, lift etc.)[1]; all of these measures of interestingness use different methods (frequency, probability and counts) to calculate frequent itemsets in data sets. The frequency of items represents basic interestingness in association rules. A main origin of measures of interestingness is from common mathematical and information theories such as found in statistics, e.g., Yule's Q method, Yule's Y method, correlation coefficient and odds ratio. Out of the 40 different measures of interestingness available in the literature, no single measure of interestingness is perfect to calculate interestingness in every ARM task. In this paper, based on

M. Song et al. (Eds.): DaWaK 2020, LNCS 12393, pp. 38–47, 2020.
https://doi.org/10.1007/978-3-030-59065-9_4

the diversity of measures proposed to date, we are identifying their roles, classifying their usefulness from several perspectives to start an extended discussion on different properties of measures of interestingness.

Issues in Selecting Measures of Interestingness in ARM

(i) A large number of measures of interestingness are available to choose and many of these measures are not useful in each ARM task.
(ii) The classical measures of interestingness generate a lot of rules, most of these rules are irrelevant and redundant in many scenarios.
(iii) Based on the meaning of measure of interestingness, it's hard to decide on the appropriate measure in a concrete decision support scenario.
(iv) Various interestingness evaluation methods seem not to be rationalized. Some literature seems to simply combine several kinds of interestingness evaluations to new kinds of measures.

This paper is structured as follows. In Sect. 2, we describe expectedness and unexpectedness with respect to the roles of different measures in ARM. Section 3 focuses on the different properties for selecting the right measures of interestingness. Section 4 presents the conclusions and future work.

2 Expectedness and Unexpectedness in ARM

A simple ARM task using classical measures for a data set containing d items potentially generates $3^d - 2^d + 1$ possible association rules and most of these association rules are expected, obvious and duplicate. Take association rules for the data items {Milk, Bread, Butter} as an example. In the association rule in Eq. (1), it can be easily understood that the association of these three items is rather obvious. In ARM, obvious or common association rules can be referred to as *expected* association rules.

$$\{Milk, Bread\} \Rightarrow \{Butter\} \tag{1}$$

The main objective of ARM is to find the interesting association rules, hidden patterns and – most importantly – unexpected association rules in the data set. The association rules generated using the following combination of {Milk, Diaper, Beer} is not as obvious andy more and creates a rather novel pattern of items; in ARM, these types of association rules can be identified as unexpected association rules:

$$\{Milk, Diaper\} \Rightarrow \{Beer\} \tag{2}$$

Based on the variety of definitions of interestingness, the interestingness of an association rule can be categorized via the following nine properties [8]: (1) conciseness, (2) coverage, (3) reliability, (4) peculiarity, (5) diversity, (6) novelty, (7) surprisingness, (8) utility and (9) actionability. Descriptions of all of these properties are summarized in Table 1. Based on these nine definitions of interestingness, the measures of interestingness can be classified into three major

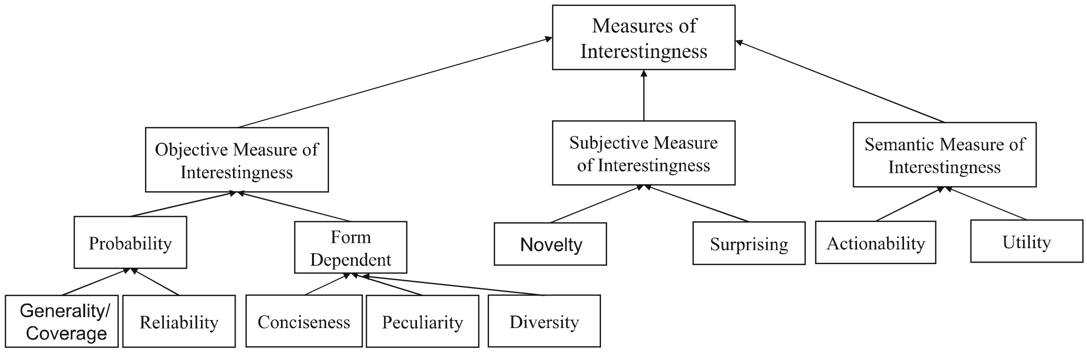

Fig. 1. Types of measures of interestingness.

categories: (1) objective measures of interestingness, (2) subjective measures of interestingness and (3) semantic measures of interestingness [14,18]. Figure 1 is showing all the different types of measures of interestingness.

2.1 Objective Measures of Interestingness for Expected Association Rules

Every transactional data set has some hidden patterns that can be easily identified by using predictive performance or statistical significance. In ARM, such kind of patterns may be referred to as expected patterns and can be computed using objective measures of interestingness. Objective measures mainly focus on the statistics and use statistical strength (probability, count etc.) to assess the degree of interest. As per the definition of interestingness, reliability, generality, conciseness, diversity and peculiarity are based only on data and patterns; therefore, these properties are the foundation of objective measures of interestingness [8]. Support, confidence, lift, conviction and improvement are some examples of objective measures of interestingness.

2.2 Subjective Measures of Interestingness for Unexpected Association Rules

Association rule mining based on common statistical approaches sometimes produces rather obvious or trivial rules. Therefore, the research of Padmanabhan and Tuzhilin [18] first explored the problem of interestingness through the notion of unexpectedness [18,19]. Subjective measures of interestingness usually determine the unexpected association rules in knowledge discovery. Unexpected patterns are opposite to the person's existing knowledge and contradict their expectations and existing knowledge [18].

Finding unexpected patterns in association rule mining is not an easy task, it needs a substantial amount of background information from domain experts [7]. For example, the association rule in Eq. (3) will rather not be considered interesting, even in cases where the rule has a particularly high support and high

Table 1. Interestingness properties in ARM, summarized and apobted from [2–6, 9, 10, 12, 15, 16, 19, 21, 23, 24, 26–28].

Property	Description
Conciseness [4, 19]	A small number of attribute-value pairs in a pattern represents the conciseness of the pattern and a set of small number of patterns refers to a concise pattern set
Generality/Coverage [2, 27]	The generality/coverage property in ARM covers most of the general patterns in ARM
Reliability [16, 24]	Association rules or patterns based on common and popular relationships can be identified as reliable association rules or patterns
Peculiarity [3, 28]	Peculiarity refers to unexpected behaviour of patterns. A pattern is said to be peculiar if it is significantly different from all other discovered patterns
Diversity [9]	For a pattern, diversity refers to the degree of differences between its elements; for a pattern set, diversity refers to the degree of differences in between the patterns
Novelty [21]	Combinations of unexpected items which create a pattern unknown to a person are known as novel patterns in ARM. These types of patterns can be discovered but can not be identified easily
Surprisingness [5, 10, 23]	Patterns which are opposite to a person's existing knowledge or expectations or create contradictions are known as surprising patterns in ARM
Utility [6, 15]	Patterns which contribute to reaching a goal are called patterns with utility. Patterns with utility allow the user to define utility functions to get particular information from data
Actionability/Applicability [12, 26]	Patterns with actionability allow a person to do a specific task for their benefits. These types of patterns usually reflect the person's action to solve a domain problem [12]

confidence, because the relationship expressed by the rule might be rather obvious to the analyst. As opposed to this, the association rule between $Milk$ and $Shaving\ Blades$ in Eq. (4) might be much more interesting, because the relationship is rather unexpected and might offer a unique opportunity for selling to the retail store.

$$\{Bread\} \Rightarrow \{Milk\} \tag{3}$$

$$\{Milk\} \Rightarrow \{Shaving\ Blades\} \tag{4}$$

Unexpectedness in Association Rule Mining. Many different definitions of unexpectedness have been proposed in the literature. In [18], unexpectedness has been defined with respect to association rules and beliefs. An association rule $P \Rightarrow Q$ is unexpected in regards to the belief $X \Rightarrow Y$ on a data set D if it follows the following rules:

- (i) $Q \wedge Y \models FALSE$ (This property states that Q and Y logically contradict each other.)
- (ii) This property states that set $P \wedge X$ has a large subset of tuples in the data set D.
- (iii) Rule $P, X \Rightarrow Q$ holds. As per the property (i), Q and Y logically contradict each other, therefore it logically follows that $P, X \Rightarrow \neg$ Y.

2.3 Semantic Measures of Interestingness

In ARM, semantic measures are a special kind of subjective measures of interestingness which include utility, application-specific semantics of patterns and domain knowledge of the person.

Utility: A utility function reflects the clear goal of the user. For example, to check the occurrence of a rare disease, a doctor might select association rules that correspond to low support rules over those with higher. A user with additional domain knowledge can use a utility-based approach. The domain knowledge of the user does not relate to his personal knowledge and expectations from data.

Actionability: In ARM, there is no widespread way to measure the actionability, i.e., it is up to the ability of an organization to do something useful with a discovered pattern; therefore, a pattern can be referred to as interesting if it is both actionable and unexpected. Generally, actionability is associated with a pattern selection strategy, whereas existing measures of interestingness are dependent on applications.

3 Properties for Selecting Objective Measures of Interestingness

It is important to care for applying consistent sets of measures of interestingness, as sometimes a wrong selection of measures may produce conflicting results. To select appropriate objective measures of interestingness, 15 key properties have been introduced in the literature [8,11,20,24]. Some of these properties are well known and some of the properties are not as popular. These properties are very useful to select appropriate measures for an ARM task.

Piatetsky-Shapiro [20] proposed three basic properties that need to be followed by every objective measure R

Property P1: "$R = 0$ if X, Y are two statistically independent data items, i.e., $\mathsf{P}(XY) = \mathsf{P}(X)\mathsf{P}(Y)$". This property states that accidentally occurred patterns or association rules are not interesting.

Property P2: "R monotonically increases with $\mathsf{P}(XY)$ when $\mathsf{P}(X)$ and $\mathsf{P}(Y)$ are same". *P2* states that if a rule $X \Rightarrow Y$ have more positive correlation then the rule is more interesting.

Property P3: "R monotonically decreases when other parameters $\mathsf{P}(X)$, $\mathsf{P}(Y)$, $\mathsf{P}(X, Y)$ remain unchanged."

Tan et al. [24] based on 2×2 contingency tables, Tan et al. [24] proposed five more properties for probability-based objective measures.

Property O1: "A measure of interestingness R is *symmetric under variable permutation* if it is preserved under the transformation $\Rightarrow_p$ of variable permutation, where $\Rightarrow_p$ is defined as matrix transpose as usual."

$$\begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow_p \begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline A & x & r \\ \hline \neg A & y & s \\ \hline \end{array}$$

Property O2: "R is same in row and column scaling. This property is known as the row-and-column scaling invariance."

$$\begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline \mathrm{A} & \mathrm{x} & \mathrm{y} \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline A & k_3k_1x & k_4k_1y \\ \hline \neg A & k_3k_2r & k_4k_2s \\ \hline \end{array}$$

Property O3: "R is anti-symmetric under row and column permutation."

$$\begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline A & r & s \\ \hline \neg A & x & y \\ \hline \end{array}$$

Property O4: "R should remain same under both row and column permutation. This is inversion invariance which shows a special case of the row/column permutation where both rows and columns are swapped simultaneously."

$$\begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline A & x & y \\ \hline \neg A & r & s \\ \hline \end{array} \Rightarrow \begin{array}{|c|c|c|} \hline & B & \neg B \\ \hline A & s & r \\ \hline \neg A & y & x \\ \hline \end{array}$$

Property O5: "This property represents the null invariance."

	B	$\neg B$
A	x	y
$\neg A$	r	s

$\Rightarrow$

	B	$\neg B$
A	x	y
$\neg A$	r	$s + k$

Lenca et al. [11] proposed five more properties to evaluate measures of interestingness. In these properties, Q1, Q4 and Q5 properties are preferred over the Q2, Q3 properties

Property Q1: "An interesting measure R is constant if there is no counterexample to the rule". As per this property all the association rules with confidence 1 should have same interestingness value.

Property Q2: "R decreases with $\mathsf{P}(X\neg Y)$ in a linear, concave, or convex fashion around 0+." This property describes that the value of interestingness decreases with respect to the counterexamples.

Property Q3: "R increases as the total number of records increases."

Property Q4: "The threshold is easy to fix." This property focuses on selecting the easy threshold to separate the interesting association rules from uninteresting association rules.

Property Q5: "The semantics of the measures are easy to express." As per this property, semantics of the interestingness measures should be understandable.

Hamilton et al. [8] have also proposed two more properties to select the right measures of interestingness.

Property S1: "An interesting measure R should be an increasing function of support if the margins in the contingency table are fixed."

Property S2: "An Interesting measure R should be an increasing function of confidence if the margins in the contingency table are fixed."

3.1 Towards Selecting Optimal Measures of Interestingness

All three categories of measures (objective, subjective and semantic) consist of many different measures; therefore, it is very difficult to select appropriate measures for an ARM task. Table 2 might be a useful step in the selection of optimal measures of interestingness.

With respect to objective measures of interestingness, Tan et al. and Lenca et al. [11,24] proposed a ranking method to select measures. The ranking method is based on a specific data set that allows specific patterns having greatest standard deviations in all of the rankings. Lenca et al. [11] proposed also another approach to select measures; in this approach, a value and a weight is assigned to each important property in purpose of selecting measures. In the approach proposed by Vaillant et al. [25], objective measures of interestingness are grouped according to their properties and outcomes.

Table 2. Suggested approaches for selecting optimal measures of interestingness.

Objective Measures of Interestingness	Subjective Measures of Interestingness	Semantic Measures of Interestingness
Ranking method based on data sets [24]	Approaches based on formal specification of user knowledge [10,13,23]	Utility-based [22]
Ranking method based on properties of measures of interestingness [11]	Eliminating uninteresting patterns [21]	Actionable patterns [13]
Clustering method based on data sets [25]	Constraining the search space [17]	–
Clustering method based on properties of measures of interestingness [25]	–	–

In subjective measures of interestingness, user knowledge and data are the two crucial factors in deciding on optimal measures. Based on existing and vague knowledge of the user, Liu et al. [13] proposed different subjective measures. The approach proposed by Sahar et al. [21] is about eliminating uninteresting patterns; in this approach, there is no specific measure of interestingness. The method proposed by Padmanabhan et al. [17] is about constraining the search space , here, user belief is used as a constraint in mining association rules. In this method, a user's belief is mined as an association rules and if existing knowledge contradicts to the mined belief, it is referred to as a surprising pattern.

With respect to selecting optimal semantic measures of interestingness, [22] have proposed an approach that is about patterns with utility, here, "Interestingness (of a pattern) = probability + utility" [22]. In the actionability approach proposed by [13], a user provides some patterns in the form of fuzzy rules to represent both the possible actions and the situations in which they are likely to be taken.

4 Conclusion

In ARM, it is clear that no single measure of interestingness is suitable for all ARM tasks – a combination of subjective measures and objective measures seem to be the future in ARM. Selecting optimal measures of interestingness is still an open research problem. In this paper, we have conducted a preliminary study of properties that have been proposed to select optimal measures of interestingness.

We have summarized the role of expected and unexpected association rules in data mining and discussed the importance of the degree of user-involvement within the ARM process. Based on this preliminary work, we aim to design a user interface that supports the decision maker in selecting optimal measures of interestingness. The findings should also be helpful in efforts of designing new measures of interestingness in the future.

Acknowledgements. This work has been conducted in the project "ICT programme" which was supported by the European Union through the European Social Fund.

References

1. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. ACM SIGMOD Record **22**(2), 207–216 (1993). https://doi.org/10.1145/170036.170072
2. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of VLDB'1994 - the 20th International Conference on Very Large Data Bases, p. 487–499. Morgan Kaufmann (1994)
3. Barnett, V., Lewis, T.: Outliers in Statistical Data. Wiley, 3rd edn (1994)
4. Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., Lakhal, L.: Mining minimal non-redundant association rules using frequent closed itemsets. In: Lloyd, J., et al. (eds.) CL 2000. LNCS (LNAI), vol. 1861, pp. 972–986. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-44957-4_65
5. Bay, S.D., Pazzani, M.J.: Detecting change in categorical data: mining contrast sets. In: Proceedings of KDD 1999 - the 5th ACM International Conference on Knowledge Discovery and Data Mining, pp. 302–306. ACM (1999)
6. Chan, R., Yang, Q., Shen, Y.D.: Mining high utility itemsets. In: Proceedings of ICDM'2003-the 3rd IEEE International Conference on Data Mining, pp. 19–26. IEEE, USA (2003)
7. Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) DEXA 2019. LNCS, vol. 11706. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-27615-7
8. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: a survey. ACM Comput. Surv. **38**(3) 56–63 (2006). https://doi.org/10.1145/1132960.1132963
9. Hilderman, R.J., Hamilton, H.J.: Knowledge Discovery and Measures of Interest. Kluwer (2001)
10. Lee, Y.C., Hong, T.P., Lin, W.Y.: Mining association rules with multiple minimum supports using maximum constraints. Int. J. Approximate Reason. **40**(1–2), 44–54 (2005). https://doi.org/10.1016/j.ijar.2004.11.006
11. Lenca, P., Meyer, P., Vaillant, B., Lallich, S.: A multicriteria decision aid for interestingness measure selection. Technical report LUSSI-TR-2004-01-EN, École Nationale Supérieure des Télécommunications de Bretagne (2004)
12. Ling, C.X., Chen, T., Yang, Q., Cheng, J.: Mining optimal actions for profitable CRM. In: Proceedings of ICDM'2002 - the 2nd IEEE International Conference on Data Mining, pp. 767–770. IEEE (2002)
13. Liu, B., Hsu, W., Chen, S.: Using general impressions to analyze discovered classification rules. In: Proceedings of KDD 1997 - The 3rd International Conference on Knowledge Discovery and Data Mining, pp. 31–36. AAAI (1997)

14. Liu, B., Hsu, W., Chen, S., Ma, Y.: Analyzing the subjective interestingness of association rules. IEEE Intell. Syst. Appl. **15**(5), 47–55 (2000). https://doi.org/10.1109/5254.889106
15. Lu, S., Hu, H., Li, F.: Mining weighted association rules. Intell. Data Anal. **5**(3), 211–225 (2001)
16. Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., Yamaguchi, T.: Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In: Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D. (eds.) PKDD 2004. LNCS (LNAI), vol. 3202, pp. 362–373. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30116-5_34
17. Padmanabhan, B., Tuzhilin, A.: A belief-driven method for discovering unexpected patterns. In: Proceedings of KDD 1998 - The 4th International Conference on Knowledge Discovery and Data Mining, pp. 94–100. AAAI (1998)
18. Padmanabhan, B., Tuzhilin, A.: Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Syst. **27**(3), 303–318 (1999). https://doi.org/10.1016/S0167-9236(99)00053-6
19. Padmanabhan, B., Tuzhilin, A.: Small is beautiful: discovering the minimal set of unexpected patterns. In: Proceedings of KDD'2000 - The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 54–63. AAAI (2000). http://doi.acm.org/10.1145/347090.347103
20. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Piatetsky-Shapiro, G., Frawley, W.J. (eds.) Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press (1991)
21. Sahar, S.: Interestingness via what is not interesting. In: Proceedings of ACM SIGKDD'1999–The 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 332–336 (1999)
22. Shen, Y.D., Zhang, Z., Yang, Q.: Objective-oriented utility-based association mining. In: Proceedings of ICDM'2002–The 2nd IEEE International Conference on Data Mining, pp. 426–433. IEEE Computer Society (2002)
23. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Trans. Knowl. Data Eng. **8**(6), 970–974 (1996). https://doi.org/10.1109/69.553165
24. Tan, P.N., Kumar, V., Srivastava, J.: Selecting the right interestingness measure for association patterns. In: Proceedings of ACM SIGKDD' 2002–The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 2, pp. 32–41 (2002). https://doi.org/10.1145/775052.775053
25. Vaillant, B., Lenca, P., Lallich, S.: A Clustering of Interestingness Measures. In: Suzuki, E., Arikawa, S. (eds.) DS 2004. LNCS (LNAI), vol. 3245, pp. 290–297. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30214-8_23
26. Jensen, C.S., et al. (eds.) EDBT 2002. LNCS, vol. 2287. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-45876-X
27. Webb, G.I., Brain, D.: Generality is predictive of prediction accuracy. In: Proceedings of PKAW'2002- Pacific Rim Knowledge Acquisition Workshop, pp. 117–130 (2002)
28. Zhong, N., Yao, Y.Y., Ohishima, M.: Peculiarity oriented multidatabase mining. IEEE Trans. Knowl. Data Eng. **15**(4), 952–960 (2003)