Information Sciences 216 (2012) 1–16

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Properties of rule interestingness measures and alternative approaches to normalization of measures

Salvatore Greco [a], Roman Słowiński [b,c,*], Izabela Szczęch [b]

[a] Department of Economics and Business, University of Catania, Corso Italia, 55, 95129 Catania, Italy
[b] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
[c] Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland

ARTICLE INFO

*Article history:*
Received 24 October 2011
Received in revised form 10 April 2012
Accepted 25 May 2012
Available online 2 June 2012

*Keywords:*
Rule interestingness measures
Properties of measures
Confirmation
Normalization

ABSTRACT

We are considering properties of interestingness measures of rules induced from data. These are: Bayesian confirmation property, two properties related to the case of entailment or refutation, called $(Ex_1)$ and logicality L, and a group of symmetry properties. We propose a modification of properties $(Ex_1)$ and L, called weak $(Ex_1)$, and weak L, that deploy the concept of confirmation in its larger sense. We demonstrate that properties $(Ex_1)$ and L do not fully reflect such understanding of the confirmation concept, and thus, we propose to substitute $(Ex_1)$ by weak $(Ex_1)$ and L by weak L. Moreover, we introduce four new approaches to normalization of confirmation measures in order to transform measures so that they would obtain desired properties. The analysis of the results of the normalizations of the confirmation measures takes into account all considered properties. We advocate for two normalized confirmation measures: measure *Z* considered in the literature, and newly proposed measure *A*. Finally, we provide some ideas for combining them in a single measure keeping all desirable properties.



## 1. Introduction

One of the main objectives of data mining process is to identify "*valid*, *novel*, *potentially useful*, and *ultimately comprehensible* knowledge from databases" [9,33]. The discovered knowledge (patterns) is often expressed in a form of "*if*..., *then*..." rules, which are consequence relations reflecting relationship, association, causation, etc., between independent (i.e. those in the premise of the rule) and dependent (i.e. those in the conclusion of the rule) attributes. The number of rules discovered in databases is often overwhelmingly large rising an urgent need to identify the most useful ones and filter out those that are irrelevant. In order to help to deal with this problem, various quantitative measures of rule interestingness (attractiveness) have been proposed and studied. Among the most commonly used interestingness measures there are *support*, *confidence*, *lift*, *rule interest function* (for a survey on interestingness measures see [3,15,25,27,31]).

Each of the measures proposed in the literature has been introduced to reflect different characteristics of rules. For example, measures like *support* [1] value generality (also referred to as coverage) of the rule, i.e. favor rules that cover a relatively large subset of a dataset. In opposition, there are measures that bring forth peculiarity, believing that patterns far away from other discovered knowledge, according to some distance measure, may be unknown to the user and therefore interesting. The list of characteristics that are emphasized by different measures is long and covers conciseness, reliability, novelty, surprisingness, utility, actionability, among others [15].

* Corresponding author at: Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland.
*E-mail addresses:* salgreco@unicit.it (S. Greco), Roman.Slowinski@cs.put.poznan.pl (R. Słowiński), Izabela.Szczech@cs.put.poznan.pl (I. Szczęch).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.05.018

Generally, interestingness measures can be categorized as *objective* and *subjective* measures. The first group can be established through statistical arguments derived from data to determine whether a rule is interesting or not. No knowledge about the user or application is needed. For example, rules that cover only very few objects from the dataset, and can therefore capture spurious relationships in data, are discarded by objective measures [21].

On the other hand, the group of subjective measures takes into account both the data and the user, thus, those measures require interaction with the user to obtain information about the user's background knowledge and expectations. Subjective measures regard a rule as uninteresting unless it reveals unexpected information about the data or provides knowledge that can lead to profitable actions [39,40]. Thus, for subjective evaluation criteria rare cases in the data are often interesting and rules that cover them are of high value.

All in all, objective measures depend on the structure of the rules and the underlying data used in the discovery process, whereas the subjective measures also rely on the class of users who examine the rule [35].

Moreover, measuring the interestingness of discovered patterns receives recently much attention from researchers developing the paradigm of granular computing (see, e.g., the rough-set-based granular computing in [2,17,18,32,37,38]).

A common conclusion stemming from this broad interest in measuring attractiveness of discovered rules is that there is no single way that would work the best on any real-life problem. The literature is a rich resource of ordinally non-equivalent measures that reflect different characteristics of rules and rank them in different ways. As there is no agreement which measure is the best, the choice of an interestingness measure for a particular application is a non-trivial task that should closely relate to the domain of application and should take advantage of available domain knowledge.

To help to analyze objective measures and to choose one for a certain application, some properties have been proposed. They express the user's expectations towards the behavior of measures in particular situations. Those expectations can be of various types, e.g., one could desire to use only such measures that reward the rules having a greater number of objects supporting the pattern. In general, properties group the measures according to similarities in their characteristics, thus using the measures which satisfy the desirable properties one can avoid considering unimportant rules. Different properties have been proposed and surveyed in [5,8,15,16,19,25,36,39].

Among the commonly used properties of rule interestingness measures there are:

- *property of confirmation* related to quantification of the degree to which the premise of the rule provides evidence for or against the conclusion [5,12];
- *property* $(Ex_1)$ assuring that any conclusively confirmatory rule is assigned a higher value of interestingness measure than any rule which is not conclusively confirmatory, and any conclusively disconfirmatory rule is assigned a lower value than any rule which is not conclusively disconfirmatory [7,20];
- *property* L, called *logicality*, for which any conclusively confirmatory rule is assigned the maximum value, and any conclusively disconfirmatory rule is assigned the minimum value [7,12]; properties $(Ex_1)$ and L can be regarded as strongly related, as both of them deal with the behavior of confirmation measures in cases of conclusive confirmation or conclusive disconfirmation;
- *properties of symmetry* being a whole set of properties that describe desirable and undesirable behavior of measures in cases when the premise or conclusion in not satisfied, or when the premise and conclusion switch positions in a rule [4,7,8,12].

This paper concentrates on the abovementioned properties of objective interestingness measures. We propose a modification of properties $(Ex_1)$ and L, called *weak* $(Ex_1)$, and *weak* L, that deploy the concept of confirmation in its larger sense. In fact, according to the deep meaning of the confirmation concept, a confirmation measure should give an account of the credibility that it is more probable to have the conclusion when the premise is present, rather than when the premise is absent. We demonstrate that properties $(Ex_1)$ and L do not fully reflect such understanding of the confirmation concept, and thus, we propose to substitute $(Ex_1)$ by weak $(Ex_1)$ and L by weak L.

Moreover, since Crupi et al. [7] represent Bayesian approach to defining $(Ex_1)$ and L, we enrich their point of view by considering also likelihoodist counterparts of those properties, denoted as L-$(Ex_1)$ and L-L, respectively.

Next, we introduce four new approaches to normalization of confirmation measures in order to transform measures so that they would obtain desired properties. The analysis of the results of the normalizations of the confirmation measures considers property $(Ex_1)$, L-$(Ex_1)$, weak $(Ex_1)$, L, L-L, weak L, and the properties of symmetry.

As the final contribution, we propose a new measure $A$ that fulfils all the desirable symmetry properties. Its strength lies in the fact that it does not possess the property $(Ex_1)$, but its likelihoodist counterpart L-$(Ex_1)$. On the basis of these remarks, we argument that measure $A$ and measure $Z$ proposed by Crupi et al. [7], should be considered as complementary tools for assessing the quality of rules. At the end, we provide some ideas for combining them in a single measure keeping all desirable properties.

## 2. Preliminaries

A *rule* induced from a dataset on a universe $U$ shall be denoted by $E \rightarrow H$ (read as "*if E, then H*"). It consists of a premise (evidence) $E$ and a conclusion (hypothesis) $H$.