FOUNDATIONS OF COMPUTING AND DECISION SCIENCES
Vol. 43 (2018) No. 1
ISSN 0867-6356
DE GRUYTER OPEN
DOI: 10.1515/fcds-2018-0004 e-ISSN 2300-3405

Can Confirmation Measures Reflect Statistically Sound Dependencies in Data? The Concordance-based Assessment

Robert Susmaga, Izabela Szczęch *

Abstract. The paper considers particular interestingness measures, called confirmation measures (also known as Bayesian confirmation measures), used for the evaluation of "*if* evidence, *then* hypothesis" rules. The agreement of such measures with a statistically sound (significant) dependency between the evidence and the hypothesis in data is thoroughly investigated. The popular confirmation measures were not defined to possess such form of agreement. However, in error-prone environments, potential lack of agreement may lead to undesired effects, e.g. when a measure indicates either strong confirmation or strong disconfirmation, while in fact there is only weak dependency between the evidence and the hypothesis. In order to detect and prevent such situations, the paper employs a coefficient allowing to assess the level of dependency between the evidence and the hypothesis in data, and introduces a method of quantifying the level of agreement (referred to as a concordance) between this coefficient and the measure being analysed. The concordance is characterized and visualised using specialized histograms, scatter-plots, etc. Moreover, risk-related interpretations of the concordance are introduced. Using a set of 12 confirmation measures, the paper presents experiments designed to establish the actual concordance as well as other useful characteristics of the measures.

Keywords: Interestingness measures, confirmation measures, statistical dependency, concordance

1. Introduction

In data mining and knowledge discovery, the discovered knowledge patterns are often expressed in the form of *if-then* rules, being consequence relations representing causation, but also correlation, association, etc., between attributes describing objects.

*Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, {rsusmaga,iszczech}@cs.put.poznan.pl

To measure the relevance and utility of the discovered rules, quantitative measures, known as interestingness (attractiveness) measures, have been proposed, studied and applied (e.g. rule (anti-)support, confidence, gain, lift, χ^2 or Fisher's coefficient) [12, 17, 18, 24, 35]. They allow a reduction in the number of rules that need to be considered by ranking them and filtering out the useless ones. Among interestingness measures, an important role is played by a group called *confirmation measures* (also referred to as Bayesian confirmation measures). Generally, they express the degree to which a rule's premise (the condition part) confirms its conclusion (the decision part) [6, 10, 11]. The discussion and analysis of various confirmation measures contained in [3, 14, 35] show the advantages of using confirmation measures for the evaluation of rules.

To group the measures according to similarities in their characteristics and behaviour and thus to help the user choose the best measure for a particular application, various measure properties have been defined. Among studied properties of confirmation measures there are: monotonicity property, Ex_1 and weak Ex_1 properties, logicality L and weak L properties, maximality/minimality, and a group of symmetry properties (for a survey refer to [3, 6, 8, 14, 15]).

Unfortunately, the property analysis becomes more complex when we assume that it is conducted upon data that may be erroneous. In practice, the existence of such possible errors is a real phenomenon and must be taken into account, so that insignificant, accidental conclusions could be eliminated [19, 38]. However, this is not always the case with the existing confirmation measures, which may indicate either weak confirmation or weak disconfirmation, while there is actually a strong dependency between the evidence and the hypothesis, or either strong confirmation or strong disconfirmation, while there is only a weak dependency [31]. To examine this aspect of the confirmation measures, the paper assesses the soundness (significance) of the dependency between the evidence and the hypothesis in experimental data, and introduces a method of quantifying the level of agreement (referred to as concordance) between this assessment and the measure being analysed. In this paper the soundness is established with the χ^2-based coefficient, while the concordance is expressed as the Pearson correlation coefficient r. Needless to say, the χ^2-r combination is only one of many possible implementations of the proposed approach, chosen here for the coefficients' well-established reliability and popularity. Nevertheless, each applied coefficient has its inherent limitations and might need adjusted for the task at hand.

As an extension and development of [32], the paper introduces analyses that involve additional characterization of the concordance and its interpretations in terms of risk. In this respect the measures are classified as (generally) risk-prone or risk-averse, which may influence their comprehension and future exploitation. Additionally, the paper shows how the measures can be modified to comply with particular user expectations regarding the risk.

The rest of the paper is organized as follows. Section 2 presents the concept of (Bayesian) confirmation, defines some popular measures and reviews the basic properties of confirmation measures proposed in the literature. The central Section 3 presents all the statistical aspects of the analyses: hazards of using the confirmation measures under observational errors, the χ^2-based coefficient that allows to assess the

level of dependency between the evidence and the hypothesis in data, the concordance between this coefficient and confirmation measures, further forms of characterizing and visualizing the concordance and its risk-related interpretations. Section 4 then provides experimental results for all the selected confirmation measures. Section 5 categorizes confirmation measures in terms of concordance and contains hints as to how the property introduced in this paper can influence the process of defining new measures. Final remarks and conclusions are contained in Section 6.

2. Related works on confirmation measures and their properties

In logic-based environments, one often deals with reality-referring sets of statements and relations. Now, while the first convey some observations about a larger reality, the second convey dependencies between those observations. The observations, also called the pieces of evidence, are basically atomic in nature, e.g. "*Socrates is a man*", while the dependencies basically consist of two parts, a premise and a conclusion, e.g. "*If x is a man, then x is mortal*" (which is another way of saying "*Every man is mortal*"). Acting together, the observations and the relations allow to infer further facts about the reality, in particular previously unidentified ones.

In knowledge discovery applications, the facts are often known (observed) but the relations are usually not known, although their existence is assumed nonetheless. The discovery is then the discovery of the relations, often done by induction, which is a process of creating patterns that are true in the universe of the analysed data.

In this paper, we consider evaluation of patterns represented in the form of rules. The starting point for such rule induction process (rule mining) is a sample of a larger reality, often represented in the form of a data table. Formally, a data table (dataset) is a pair $S = (U, A)$, where U is a non-empty finite set of objects called the *universe*, and A is a non-empty finite set of *attributes* describing the object.

A rule induced from a dataset S on a universe U consists of a *premise* "*if* E" (referring to an existing piece of evidence, E), and a *conclusion* "*then* H" (referring to a hypothesised piece of evidence, H). Further on, we shall use the common, shortened denotation $E \to H$ (read as "*if E, then H*").

Typically, the number of patterns induced from datasets is quite large. It can be overwhelming for an expert analysing the data. Therefore, there is a need to measure the relevance and the utility of the discovered patterns and to filter out those that are irrelevant, misleading, or do not provide new knowledge. In particular, to evaluate the induced rules, quantitative interestingness measures have been proposed. The literature provides a wide spectrum of ordinally non-equivalent interestingness measures, exploiting different characteristics of the mined rules (see, for example, [3, 12, 24] for exhaustive reviews of the subject).

This paper concentrates on a group of interestingness measures called confirmation measures. According to Fitelson [11], such measures quantify the degree to which the evidence in the rule's premise E provides support *for* or *against* the hypothesised piece of evidence in the rule's conclusion H. Thus, the definition of Bayesian confirmation

measures is based on consideration of four particular situations: the rule's premise E holds and so does its conclusion H, the rule's premise E does not hold, but its conclusion H holds, the rule's premise E holds, but its conclusion H does not hold, and finally neither the rule's premise E nor its conclusion H holds. As far as the real-world phenomena described by the rules are concerned, specific information on whether a given piece of evidence E or hypothesis H holds or not, is often represented with binary (and thus discrete) variables: a variable reflecting the presence or absence of the evidence (denoted as VE) and a variable reflecting the presence or absence of the hypothesis (denoted as VH). In the context of a particular dataset U, these variables may be collectively processed to reveal four non-negative integers: a, b, c and d, represented in the contingency table (see Table 1), with rows and columns characterizing the premise and the conclusion, respectively. As an illustration, let us recall a popular folk statement that "*all ravens are black*", formalized as a rule "*if x is a raven, then x is black*", often used by Hempel [20]. With this rule: a is the number of black ravens, b is the number of black non-ravens, c is the number of non-black ravens, d is the number of non-black non-ravens. The notation based on values a, b, c and d shall be used throughout the paper.

Table 1. A contingency table characterization (with header and sum rows/columns) of the rule's premise and conclusion

	H	$\neg H$	Σ
E	a	c	$a+c$
$\neg E$	b	d	$b+d$
Σ	$a+b$	$c+d$	n

Reasoning in terms of a, b, c and d is natural and intuitive for data mining techniques, since all observations are collected in some kind of an information table, describing each object by a set of attributes. However, a, b, c and d can also be used to estimate probabilities: e.g. the probability of the premise is expressed as $P(E) = (a+c)/n$, and the probability of the conclusion as $P(H) = (a+b)/n$. Fortunately, $a+b+c+d = n > 0$, since $n = |U|$ and $U \neq \emptyset$, so both of the above probabilities are always defined. Additionally, the conditional probability of the conclusion given the premise is $P(H|E) = P(H \cap E)/P(E) = a/(a+c)$, which, however, is only defined when $a + c > 0$.

The group of confirmation measures, which we shall present and analyse, consists of interestingness measures that satisfy the property of Bayesian confirmation. Formally, for a rule $E \rightarrow H$, an interestingness measure $c(H, E)$ has the property of Bayesian confirmation when it satisfies the following (1) conditions (further referred to as the BC conditions):

$$c(H,E) \begin{cases} > 0 & \text{when} \quad P(H|E) > P(H), \quad \text{equivalent to} \quad \frac{a}{a+c} > \frac{a+b}{n}, \\ = 0 & \text{when} \quad P(H|E) = P(H), \quad \text{equivalent to} \quad \frac{a}{a+c} = \frac{a+b}{n}, \\ < 0 & \text{when} \quad P(H|E) < P(H), \quad \text{equivalent to} \quad \frac{a}{a+c} < \frac{a+b}{n}. \end{cases} \tag{1}$$

Regarding the BC conditions, the confirmation is interpreted as an increase in the

probability of the conclusion H provided by the premise E. Analogously, the neutrality is regarded as the lack of influence of the premise on the probability of the conclusion, and the disconfirmation as the decrease of the probability of the conclusion imposed by the premise.
Let us stress that the list of alternative, non-equivalent measures of confirmation is quite large [6, 10]. It is due to the fact that the BC conditions do not impose any constraints on the measures except for requiring when the measures should obtain positive/negative values or zero. Thus, the property of Bayesian confirmation does not favour one single measure as the most adequate. The commonly used confirmation measures are presented in Table 2.

The definitions of confirmation measures are often formulated for two of the main defined situations: that of confirmation and that of disconfirmation. While confirmation means $P(H|E) > P(H)$, disconfirmation means $P(H|E) < P(H)$. In the third possible defined situation, i.e. when $P(H|E) = P(H)$, the measures default to 0 (for reasons of brevity this situation was omitted from the definitions of measures provided in Table 2). As opposed to defined situations, undefined ones occur when the probability $P(H|E) = P(H \cap E)/P(E)$ is not defined, which, in turn, occurs when $P(E)$ is zero. Measures of conditional definitions (i.e. definitions with separate formulae in cases of confirmation and disconfirmation, e.g. $Z(H, E)$) will always be assumed to equal zero if neither confirmation nor disconfirmation holds.

For the sake of convenience and comparability of the presented results, the greatest lower bound (infimum) and the least upper bound (supremum) of all of the analysed measures have been unified to -1 and $+1$, respectively. While most of the definitions of the considered measures could be used directly, the definition of measure $C(H, E)$ required a simple transformation. Measure $C(H, E)$ originally obtains values from $-1/4$ to $+1/4$ (regardless of n), thus all further results are presented using the rescaled $C(H, E)$. Measures $Z(H, E)$ and $A(H, E)$ are defined using a real-valued parameter p, where $p > 0$. This parameter influences what is referred to as curvature of the measure. The (default) value of 1 is assumed for p throughout this paper. Measures $c_1(H, E)$ and $c_2(H, E)$ are defined using parameters α and β, where $\alpha + \beta = 1$ and $\alpha > 0$, $\beta > 0$. Observe that parameters α and β can be used to bring the new measure closer to $Z(H, E)$ or $A(H, E)$ (defined for $p = 1$). The (default) value of 0.5 is assumed for both α and β throughout this paper.

Below, let us recall the references of the 12 analysed confirmation measures. Measure $D(H, E)$ has been considered among others by Earman [7]. Measure $M(H, E)$ has been supported by Mortimer [26]. Measure $S(H, E)$ has been proposed by Christensen [5] and Joyce [21]. Measure $N(H, E)$ has been considered by Nozick [27] and measure $C(H, E)$ by Carnap [4]. Measure $F(H, E)$ has been supported by Kemeny and Oppenheim [23] and Pearl [28]. Fitelson [10, 11] has also advocated for measure $F(H, E)$. Measure $Z(H, E)$ has been recently introduced by Crupi, Tentori and Gonzalez [6] as a measure resulting from a transformation of measures $D(H, E)$, $M(H, E)$, $S(H, E)$, $N(H, E)$ or $C(H, E)$, and possessing some valuable properties concerning situations when a rule's premise entails or refutes its conclusion. Measure $A(H, E)$ has been presented by Greco, Słowiński and Szczęch [15] as a likelihoodist counterpart of measure $Z(H, E)$, complementing the properties of measure $Z(H, E)$. Finally,

Table 2. Popular confirmation measures

$D(H,E) = P(H\|E) - P(H) = \dfrac{a}{a+c} - \dfrac{a+b}{n}$
$M(H,E) = P(E\|H) - P(E) = \dfrac{a}{a+b} - \dfrac{a+c}{n}$
$S(H,E) = P(H\|E) - P(H\|\neg E) = \dfrac{a}{a+c} - \dfrac{b}{b+d}$
$N(H,E) = P(E\|H) - P(E\|\neg H) = \dfrac{a}{a+b} - \dfrac{c}{c+d}$
$C(H,E) = P(E \wedge H) - P(E)P(H) = \dfrac{a}{n} - \dfrac{(a+c)(a+b)}{n^2}$
$F(H,E) = \dfrac{P(E\|H) - P(E\|\neg H)}{P(E\|H) + P(E\|\neg H)} = \dfrac{ad-bc}{ad+bc+2ac}$
$Z(H,E) = \begin{cases} 1 - \frac{P(\neg H\|E)}{P(\neg H)} = \frac{ad-bc}{(a+c)(c+d)} & \text{in case of confirmation} \\ \frac{P(H\|E)}{P(H)} - 1 = \frac{ad-bc}{(a+c)(a+b)} & \text{in case of disconfirmation} \end{cases}$
$A(H,E) = \begin{cases} \frac{P(E\|H)-P(E)}{1-P(E)} = \frac{ad-bc}{(a+b)(b+d)} & \text{in case of confirmation} \\ \frac{P(H)-P(H\|\neg E)}{1-P(H)} = \frac{ad-bc}{(b+d)(c+d)} & \text{in case of disconfirmation} \end{cases}$
$c_1(H,E) = \begin{cases} \alpha + \beta A(H,E) & \text{in case of confirmation when } c = 0 \\ \alpha Z(H,E) & \text{in case of confirmation when } c > 0 \\ \alpha Z(H,E) & \text{in case of disconfirmation when } a > 0 \\ -\alpha + \beta A(H,E) & \text{in case of disconfirmation when } a = 0 \end{cases}$
$c_2(H,E) = \begin{cases} \alpha + \beta Z(H,E) & \text{in case of confirmation when } b = 0 \\ \alpha A(H,E) & \text{in case of confirmation when } b > 0 \\ \alpha A(H,E) & \text{in case of disconfirmation when } d > 0 \\ -\alpha + \beta Z(H,E) & \text{in case of disconfirmation when } d = 0 \end{cases}$
$c_3(H,E) = \begin{cases} A(H,E)Z(H,E) & \text{in case of confirmation} \\ -A(H,E)Z(H,E) & \text{in case of disconfirmation} \end{cases}$
$c_4(H,E) = \begin{cases} min(A(H,E), Z(H,E)) & \text{in case of confirmation} \\ max(A(H,E), Z(H,E)) & \text{in case of disconfirmation} \end{cases}$

measures $c_1(H,E)$ - $c_4(H,E)$ have been proposed by Greco, Słowiński and Szczęch [15] as measures derived from $Z(H,E)$ and $A(H,E)$ in such way that they satisfy

desirable properties.

Choosing a measure for a particular application is usually a difficult task. There is no indication as to which measure is generally the best, as each of them captures different characteristics of the data. To group the measures according to similarities in their behaviour, and this way help to choose an appropriate measure in a particular situation, various *properties* of measures have been proposed and studied. Analysis of measures with respect to their properties is an important research area, because using the measures that satisfy the desirable properties one can avoid unimportant rules [14, 35]. Different properties have been proposed and surveyed in [4, 8, 12, 24]. Among the commonly used properties of confirmation measures there are such properties as:

- *property* M, ensuring monotonic dependency of the measure on the number of objects satisfying (supporting) or not the premise and/or the conclusion of the rule [3, 14, 35];
- *property* Ex_1 (and its modification to *weak* Ex_1), assuring that any conclusively confirmatory rule is assigned a higher value of a 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 [6, 15];
- *logicality* L (and its modification to *weak* L), indicating the conditions under which measures should obtain their maximal or minimal values [6, 11, 15];
- *maximality/minimality property* requiring that measures should obtain their maximal (or minimal) values if and only if $c = b = 0$ ($a = d = 0$) [13];
- *properties of symmetry* being a whole set of properties that describe desirable and undesirable behaviour of measures in cases when the premise or conclusion in not satisfied, or when the premise and conclusion switch positions in a rule [6, 8, 16, 34].

The considerations about properties of confirmation measures reveal the lack of a desirable property that would prevent measures from "presenting radical opinions" (i.e. obtaining extreme or close to extreme values) on the basis of data that could be statistically unsound or that could represent observational errors.

3. Using confirmation measures under observational errors

Let us return to the example regarding the black ravens, with the statement (rule) "*if x is a raven, then x is black*" decomposed into "*x is a raven*" (the evidence, E) and "*x is black*" (the hypothesis, H). To verify this rule empirically, we would look for x's that can be classified as being ravens (E) and being non-ravens ($\neg E$) on one hand, and as being black (H) and being not black ($\neg H$) on the other. These classified observations can be expressed in terms of the a, b, c and d frequencies, as described in Section 2.

Assume $a+c>0$, $b+d>0$, $a+b>0$ and $c+d>0$, which means that all possible situations, e.g. E, $\neg E$, H and $\neg H$, have occurred. According to one interpretation of such data, $c=0$ implies that the rule described with them can be treated as a proper implication (notice the similarity between the corresponding contingency table and the "truth table" of the logical function of implication), while any $c>0$ means that the contingency table represents a relation that is not an implication (although approximate, or fuzzy, implications could also be considered and applied, [37]). In such circumstances, as long as no non-black raven is observed, i.e. as long as $c=0$, the rule "*all ravens are black*" holds. However, it gets conclusively invalidated by any single non-black raven, i.e. after having observed at least one non-black raven ($c \geq 1$), this rule does not hold.

The analysis becomes more complex when we assume that the process of making observations may be error-prone. In those circumstances, having observed a non-black raven does not conclusively invalidate the rule (neither does it make it approximate), instead, it may merely imply that the observation was erroneous.

In real-life situations the existence of such possible errors must be taken into account, as potentially false conclusions might be drawn otherwise. This assumption also calls for some additional ones, in particular, the assumption about the errors being not systematic (which means that they do not concern all the observations), but being relatively rare and random. Moreover, all types of errors (e.g. registering a non-raven as a raven) are assumed to occur equally often.

After having assumed the possibility of making observational errors, a single non-black raven should not conclusively invalidate the "*all ravens are black*" rule, because an observational error may be involved. But this is not to say that every possible number of non-black ravens could be attributed to observational errors, and that such a rule can therefore never be invalidated. Conclusive invalidations of rules are certainly possible, but they require that the number of the observations against a given rule becomes, in a sense, large enough in relation to all other observations. This, of course, means that the process of invalidating a given rule amounts to determining whether the number of observations against this rule is actually large enough in relation to all other observations.

Articulating error sensitivity may be endeavoured with a tool designed to test for the independency of two discrete-valued variables. This is because while measures of confirmation and those of independency between variables measure formally different things, confirmation and independency are certainly not completely disconnected notions, especially with discrete variables of low-cardinality domains. This may be illustrated as follows.

Let VE and VH be binary variables (with $\{0,1\}$ domains), E be equivalent to $VE=1$, and H be equivalent to $VH=1$. In this case, $E \implies H$ is expressed as $VE=1 \implies VH=1$. However, $(E \implies H) \iff (\neg H \implies \neg E)$, also expressed as $(VE=1 \implies VH=1) \iff (VH=0 \implies VE=0)$. Notice that when VE and VH are fully dependent, the support of $VE=1 \implies VH=1$ and $VH=0 \implies VE=0$ (and thus of $E \implies H$ and $\neg H \implies \neg E$) will attain its maximum. This is why (as attempted in this paper) the behaviour of confirmation measures may be approached similarly to that of the dependency (or independency) measures (well established in

error-prone environments).

One of the most popular tests for the independency of two discrete-valued variables is the two-dimensional χ^2 test. The test derives its name from the χ^2-distributed coefficient, which allows to test for statistical significance at a pre-defined significance level α_0[1]. Together with its numerous corrections and improvements, χ^2-distributed coefficient has given rise to many significance testing procedures. The popular alternatives include the Cramer's V coefficient, the Yule's Q coefficient, the Fisher's coefficient or the Pearson's Φ (also referred to as the Yule's Φ) coefficient ([1, 9, 29]).

Given a 2×2-sized contingency table $\left[\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right]$, where $a + b + c + d = n$, the χ^2_0 coefficient is defined as $\chi^2_0 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$. The values of this coefficient belong to the interval $[0, n]$, and are thus n-dependent. To make them n-independent, χ^2_0 is scaled down (divided) by n, producing a value belonging to the interval $[0, 1]$. This version of the coefficient, formally defined as $\chi^2_{01} = \frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$, will be further referred to as the "scaled-down χ^2_0" (notice that χ^2_{01} is the square of the Pearson's Φ coefficient). In case of both coefficients, values close to minimum/maximum indicate weak/strong statistical dependency between the evidence and the hypothesis.

3.1. The property of concordance C

Unrestricted application of a confirmation measure $c(H, E)$ to a contingency table is perfectly acceptable when this is free from observational errors. Unfortunately, this might not be the case in all circumstances.

Two potentially unfavourable situations that can concern the confirmation measure applied to a contingency table created from error-stricken data are as follows:

- the value of $c(H, E)$ indicates either weak confirmation or weak disconfirmation, while in fact there is a strong dependency between the evidence and the hypothesis,
- the value of $c(H, E)$ indicates either strong confirmation or strong disconfirmation, while there is only a weak dependency between the evidence and the hypothesis.

The first situation is rather unlikely, as strong dependency between the evidence and the hypothesis usually implies either strong confirmation or strong disconfirmation. It is also rather 'safe', in the sense that weak confirmation or weak disconfirmation does not usually entail further user actions.

The much worse second situation can easily occur when the number of observations is small, because then even fairly few observational errors can turn out to be quite numerous in relation to all other observations. It it also 'unsafe', in the sense that strong confirmation or strong disconfirmation might entail some strong inferences, which, in turn, might provoke some irreversible (but unjustified, and thus inappropriate) user

[1]This α_0 should not be confused with the α used in the definitions of the measures $c_1(H, E)$ and $c_2(H, E)$.

actions. Finally, the situation may occur more frequently in practice, since according to a popular "rule of thumb" high values of a measure are often treated by their users with more confidence then low ones [30].

To counteract those unfavourable situations, there arises a need to evaluate the potential concordance between confirmation measures and statistical significance of the evidence-hypothesis dependency. For such an evaluation to be useful, it should provide continuous measurements, the higher the more the measure $c(H, E)$ 'agrees' with the level of dependency between the evidence and the hypothesis. Highly 'agreeable' measures become then autonomous in the sense that high values of the measures translate to high levels of the above-mentioned dependency, and no additional testing is required.

More formally, a confirmation measure $c(H, E)$ is characterized by:

- high values of concordance C when extreme values of the confirmation measure occur only when there is a strong dependency between E and H and neutral values of the confirmation measure occur only when there is a weak statistical dependency between E and H,
- low values of concordance C otherwise.

Measures characterized by the desired, high values of concordance will be referred to as concordant measures.

Let us remark that the evaluation of the concordance may be performed using different statistical tools (e.g. correlation). Nevertheless, the desirable behaviour of the measure being analysed is as follows:

- $|c(H, E)|$ is close to 0 whenever the evidence and the hypothesis are roughly independent,
- $|c(H, E)|$ grows with the growing dependency between the evidence and the hypothesis,
- $|c(H, E)|$ is close to 1 whenever the evidence and the hypothesis are strongly dependent.

By $|c(H, E)|$ we denote the absolute value of $c(H, E)$. Taking $|c(H, E)|$ into account (and thus ignoring its sign) is essential, as it is the absolute value of the confirmation measure, and not the sign, that determines the 'strength' of $c(H, E)$ (i.e. the degree to which the premise of a rule evaluated by the confirmation measure confirms or disconfirms its conclusion).

Let us remark that the above-mentioned dependencies may be in practice quantified by any procedure capable of testing for the independency of two discrete variables, provided it is applied to the two observation-describing (binary) variables: *VE* and *VH*. A very common procedure of such type is the χ^2-based testing for the independency of two discrete variables, which may, in particular, be carried out using the χ^2_0 coefficient or its scaled-down version, χ^2_{01}. In this context, the concordance between a confirmation measure and χ^2_{01} will be referred to as χ^2_{01}-concordance and the concordant measures will be referred to as χ^2_{01}-concordant.

Under the assumption of expressing the dependency between the *VE* and *VH* with the χ^2_{01} coefficient, the relation between this coefficient and a given confirmation measure $c(H,E)$ may be additionally visualized. This is easily done with a scatter-plot of $c(H,E)$ against χ^2_{01}. Each such scatter-plot will fit a 2×1-sized rectangular envelope, with its axes ranging from -1 to $+1$ (horizontal, $c(H,E)$) and from 0 to 1 (vertical, χ^2_{01}), as illustrated in Figure 1, which features three colours: red, green and blue as well as the graded transitions (reddish-greenish and greenish-blueish) between them[2].

Given a χ^2_{01}-concordant measure $c(H,E)$, the points of the $c(H,E)$-versus-χ^2_{01} scatter-plot should possibly occupy the greenish region of the figure, while possibly avoiding both reddish and blueish regions. This is because points located in the greenish region illustrate good concordance between $c(H,E)$ and χ^2_{01}, while points located in the two other regions indicate poor χ^2_{01}-concordance. In particular, the (two distinct) reddish regions (marginal lower parts of the envelope) characterize situations in which the value of $|c(H,E)|$ is disproportionally large as compared to the value of χ^2_{01}. This means that the measure actually presents "radical opinions", despite insignificant dependency between the evidence and the hypothesis. On the other hand, the (two merged) blueish regions (central upper part of the envelope) characterize situations in which the value of $|c(H,E)|$ is disproportionally small as compared to the value of χ^2_{01}. This means that the measure actually presents "inhibited opinions", despite significant dependency between the evidence and the hypothesis.

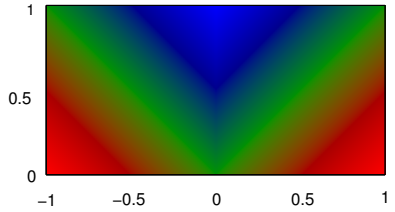

Figure 1. The desirable (greenish) and undesirable (reddish, blueish) regions of the $c(H,E)$-versus-χ^2_{01} scatter-plot

What is specific about the property of concordance C is that it is a representative of continuous-type properties: it can be quantified as the level of dependency between two continuous entities, provided the quantification is properly applied to a given $c(H,E)$ (the first continuous entity) and the measure of dependency between *VE* and *VH* (the second continuous entity). In particular, if we assume that the dependency between *VE* and *VH* is expressed with the χ^2_{01} coefficient, the dependency between this coefficient and the confirmation measure $c(H,E)$ may be represented, e.g., as the linear Pearson correlation[3], denoted as r, between χ^2_{01} and the $|c(H,E)|$. If the measure $c(H,E)$ is to be χ^2_{01}-concordant, then $|c(H,E)|$ should be strongly, positively

[2]Owing to the journal's printing policy, the version of the paper with colour rendering of the figures is available only at the journal's online web pages.

[3]Given two variables (in practice: two vectors), $\mathbf{x}$ and $\mathbf{y}$, the Pearson correlation coefficient r can be expressed as $r(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{x}||\cdot||\mathbf{y}||}$, with $||\mathbf{z}||$ denoting the (euclidean) norm of the vector $\mathbf{z}$.

correlated with χ^2_{01}, otherwise $|c(H,E)|$ can be uncorrelated to, or even negatively correlated with χ^2_{01}.

The linear Pearson correlation is only one possible way of representing the dependency between χ^2_{01} and $|c(H,E)|$, one that will be used throughout this paper[4]. Further possibilities include other, potentially non-linear correlation measures, as well as dependency measures, like principal curve based measures [2], mutual information based measures [1], etc. Also the χ^2_0-based way of expressing the dependency between VE and VH is not the only possibility. Its alternatives include the corrected χ^2 coefficient or the Fisher's test-based coefficient.

In order to provide additional characterization of the dependency between χ^2_{01} and $|c(H,E)|$, we shall also use the Spearman rank correlation coefficient[5], denoted as s, similarly applied to χ^2_{01} and $|c(H,E)|$. The important thing is that the coefficients r and s differ in that s ignores the particular differences between the compared vectors and obtains maximal value (i.e. 1) as long as these vectors are equivalent *ordinally*. In particular, while r is not sensitive to any linear (but strictly positive) transformations of the vectors, which means that given a positive value a, the same value of r is produced for $\mathbf{x}$ and $\mathbf{y}$ as well as for $\mathbf{x}$ and $\mathbf{z}$, where $z_i = ay_i$, s is not sensitive to any non-linear (but strictly monotone) transformations of the vectors, e.g. given a strictly monotone function $f(x)$, the same value of s is produced for $\mathbf{x}$ and $\mathbf{y}$ as well as for $\mathbf{x}$ and $\mathbf{z}$, where $z_i = f(y_i)$. This ordinal independency may actually prove advantageous in experiments, in which the confirmation measures are used in machine learning algorithms.

The presented property of χ^2_{01}-concordance (as quantified by the Pearson correlation coefficient) is different from and functionally independent of the previously discussed (see Section 2) properties of the confirmation measures. In particular, it is neither directly implied by nor directly implies any of those properties.

3.2. Risk-related interpretation of the property of concordance C

The character of a given dependency, especially a functional one, may allow for some very specialized inferences about this dependency [22, 25]. Given a criterion, a function referred to as the utility function expresses functional dependency between the values of this criterion (the function's domain) and what is called the utility of these values (the function's value set), as viewed by a particular decision maker. The character of the utility function $u(x)$ created by the decision maker in an iterative process may be assumed to reveal the risk attitude of this decision maker. Three basic forms of the decision maker-specific risk attitude, or decision maker profiles, are often distinguished in this case (potentially for some subintervals of the function's domain):

- the utility function $u(x)$ is convex: the decision maker is 'risk-prone',

[4] As far as the confirmation contexts is concerned, the Pearson coefficient has been used to assess similarity between different measures in [36].

[5] Given two variables (in practice: two vectors), $\mathbf{x}$ and $\mathbf{y}$, the Spearman correlation coefficient s can be expressed as $s(\mathbf{x},\mathbf{y}) = r(R(\mathbf{x}),R(\mathbf{y}))$, where $r(\mathbf{x},\mathbf{y})$ is the linear Pearson correlation, while $R(\mathbf{z})$ denotes a vector of ranks of the elements of vector $\mathbf{z}$.

- the utility function $u(x)$ is linear: the decision maker is 'risk-neutral',
- the utility function $u(x)$ is concave: the decision maker is 'risk-averse'.

Essentially similar reasoning may be applied to the $c(H,E)$-versus-χ^2_{01} dependency. In case when this dependency exists in a functional form, i.e. when there exists a function $f : [-1,+1] \to [0,1]$ satisfying $f(c(H,E)) = \chi^2_{01}$, the following measure profiles can be distinguished (potentially for some subintervals of the confirmation measure's domain):

- the function $f(x)$ is convex: the measure $c(H,E)$ is 'risk-prone',
- the function $f(x)$ is linear: the measure $c(H,E)$ is 'risk-neutral',
- the function $f(x)$ is concave: the measure $c(H,E)$ is 'risk-averse'.

In many practical situations no functional dependency between $c(H,E)$ and χ^2_{01} will exist. In such cases the approximate dependency may be visualized with scatter-plots (as introduced in subsection 3.1), and the profiles established by observing whether the points of the scatter-plot are situated below ('risk-prone') or above ('risk-averse') of the $|c(H,E)| = \chi^2_{01}$ line. In this context, the following situations may be distinguished (potentially for some subintervals of the confirmation measure's domain):

- all points of the $c(H,E)$-versus-χ^2_{01} scatter-plot are situated below the $|c(H,E)| = \chi^2_{01}$ line: the measure $c(H,E)$ is 'risk-prone',
- all points of the $c(H,E)$-versus-χ^2_{01} scatter-plot are situated on the $|c(H,E)| = \chi^2_{01}$ line: the measure $c(H,E)$ is 'risk-neutral',
- all points of the $c(H,E)$-versus-χ^2_{01} scatter-plot are situated above the $|c(H,E)| = \chi^2_{01}$ line: the measure $c(H,E)$ is 'risk-averse'.

If most, but not all points of the scatter-plot are located below, on, or above the $|c(H,E)| = \chi^2_{01}$ line, the measures are referred to as predominantly 'risk-prone', predominantly 'risk-neutral' or predominantly 'risk-averse', respectively. Now, it may be especially interesting as well as useful to provide some aggregated forms of the information conveyed by the scatter-plots. Counting points of the $c(H,E)$-versus-χ^2_{01} scatter-plot constitutes only one possible kind of such an aggregation, in which it is the numbers of points that lie below or above the $|c(H,E)| = \chi^2_{01}$ line that are actually taken into account, while the distances between this line and the given points are ignored.

4. Evaluating the concordance of confirmation measures

The following experiments have been carried out to evaluate how the 12 selected Bayesian confirmation measures reflect statistically significant dependencies[6].

[6]Source code files (MatLab scripts) are available at `http://www.cs.put.poznan.pl/iszczech/publications/MatLab_scripts_FCDS_2018.zip`.

The experimental data generally consist of a set of contingency tables, each of which contains four integer entries: a, b, c and d, $a+b+c+d=n$, as introduced in Section 2. Given $n>0$ (the total number of observations), the dataset is generated as the set of all possible $\left[\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right]$ contingency tables satisfying $a+b+c+d=n$. The set is thus exhaustive and non-redundant, as it contains exactly one copy of each contingency table satisfying $a+b+c+d=n$.

The exact number of tables t in the set is given by $t=(n+1)(n+2)(n+3)/6$ (thus, $O(n^3)$). These values grow quickly (although polynomially, not exponentially). Unfortunately, the number t can become considerable: for n about 1000 (a typical number of objects in a benchmark classification dataset) t exceeds hundreds of millions. At the same time, the exhaustive and non-redundant set of contingency tables has the decisive advantage of being essentially regular. As such, it covers the unit simplex fairly uniformly (see [31, 33]). Consequently, the data and the results produced from them may successfully be interpreted in probabilistic terms.

4.1. The experimental set-up

After having set the total number of observations n, the following operations were performed:

- the exhaustive and non-redundant set of $\left[\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right]$ contingency tables satisfying $a+b+c+d=n$ was generated (there exist exactly $t=22,632,705$ such tables for $n=512$),
- the values of the 12 selected confirmation measures $c(H,E)$: $D(H,E)$, $M(H,E)$, $S(H,E)$, $N(H,E)$, $C(H,E)$, $F(H,E)$, $Z(H,E)$, $A(H,E)$, $c_1(H,E)$, $c_2(H,E)$, $c_3(H,E)$, $c_4(H,E)$ for all the generated tables were computed,
- the values of the χ^2_{01} coefficient for all the generated tables were computed.

Thereupon the following coefficients were computed (see Table 3):

- the linear Pearson correlation coefficient $r(|c(H,E)|,\chi^2_{01})$,
- the Spearman rank correlation coefficient $s(|c(H,E)|,\chi^2_{01})$.

The values of coefficients change with n. Fortunately, as n grows infinitely, they converge, and the differences between their values generated for n and $n+1$ actually become negligible after n exceeds 512, assumed to produce the values quoted in Table 3. The actual convergence of the linear Pearson correlation coefficient $r(|c(H,E)|,\chi^2_{01})$ and the Spearman rank correlation coefficient $s(|c(H,E)|,\chi^2_{01})$ is illustrated in Figures 2 and 3.

By definition, the resulting correlation coefficients characterize the relation between any given measure and χ^2_{01} only in an aggregated, i.e. scalarized (and thus simplified) way. The more complex nature of these relations may be conveyed by

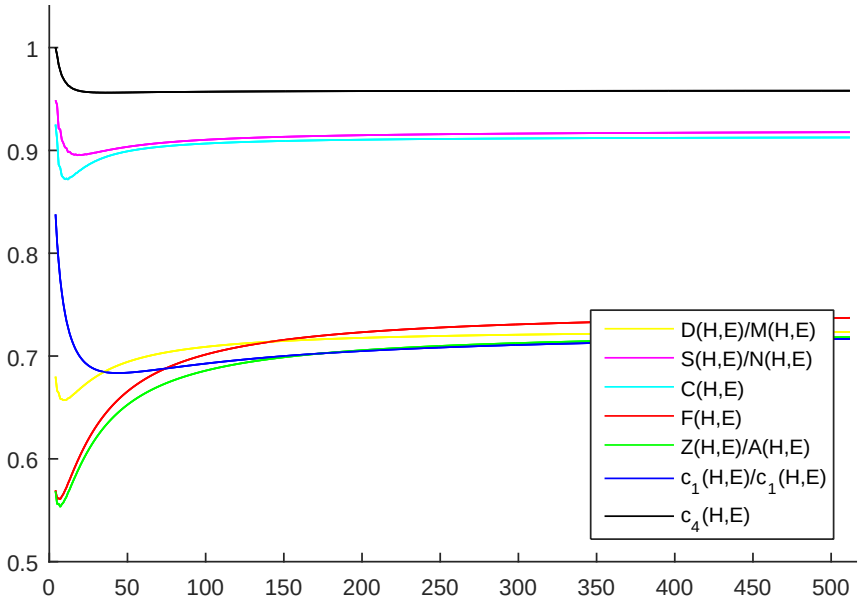

Figure 2. The convergence of $r(|c(H,E)|, \chi^2_{01})$ for $n = 4..512$

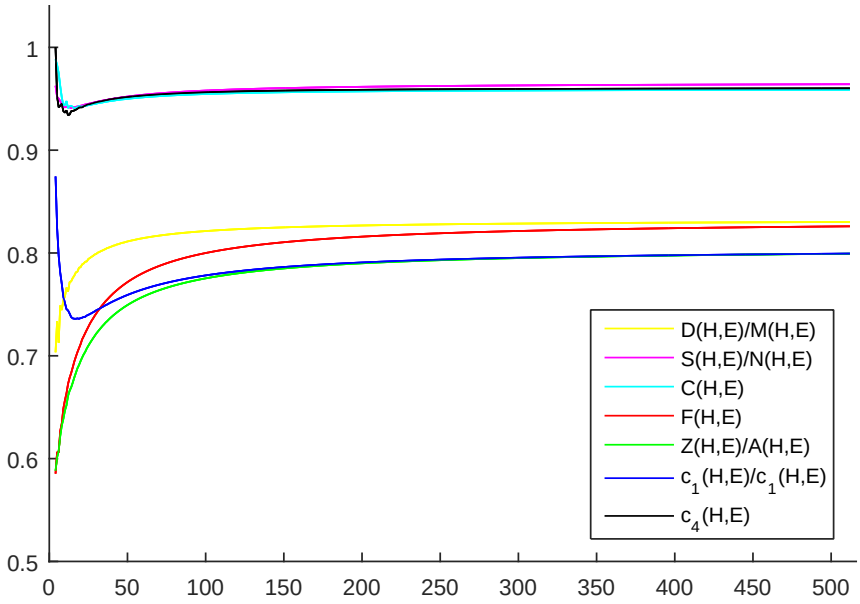

Figure 3. The convergence of $s(|c(H,E)|, \chi^2_{01})$ for $n = 4..512$

Table 3. The coefficients characterizing the χ^2_{01}-concordance of the 12 selected confirmation measures

$c(H,E)$	$r(\lvert c(H,E)\rvert, \chi^2_{01})$	$s(\lvert c(H,E)\rvert, \chi^2_{01})$
$D(H,E)$	0.723	0.830
$M(H,E)$	0.723	0.830
$S(H,E)$	0.918	0.964
$N(H,E)$	0.918	0.964
$C(H,E)$	0.913	0.958
$F(H,E)$	0.737	0.826
$Z(H,E)$	0.718	0.799
$A(H,E)$	0.718	0.799
$c_1(H,E)$	0.717	0.799
$c_2(H,E)$	0.717	0.799
$c_3(H,E)$	1.000	1.000
$c_4(H,E)$	0.958	0.960

other tools, e.g. by scatter-plots or specialized histograms. For each of the 12 selected confirmation measures the following graphical tools[7] were deployed:

- a triple-colour scatter-plot of $c(H,E)$ against χ^2_{01}, with the colours: red, green and blue (and graded transitions between them), corresponding to the points situated below, on, or above the $|c(H,E)| = \chi^2_{01}$ line, respectively (see Figure 4),
- a triple-colour histogram of $c(H,E)$, with the colours: red, green and blue, corresponding to situations in which the values of $|c(H,E)|$ are higher, equal or lower than the values of χ^2_{01}, respectively (see Figure 5).

4.2. The experimental results

The conducted experiments revealed results of both generic and specific nature, by which we mean results concerning all the 12 selected confirmation measures or results concerning only particular measures, respectively.

The following remarks concern the χ^2_{01}-concordance (as quantified by the Pearson correlation coefficient r) of the measures (see Table 3):

- the measure $c_3(H,E)$ enjoys an ideal χ^2_{01}-concordance, which is due to the fact that $|c_3(H,E)| = \chi^2_{01}$,
- the concordances of the other measures range from 0.957 (measure $c_4(H,E)$) down to 0.694 (measures $Z(H,E)$ and $A(H,E)$), in result of which all of them can be referred to as approximately concordant,

[7] Owing to the journal's printing policy, the version of the paper with colour rendering of the figures is available only at the journal's online web pages. Independently of that, $n = 32$ was always assumed when rendering the scatter-plots to reduce the size of the resulting graphics files.

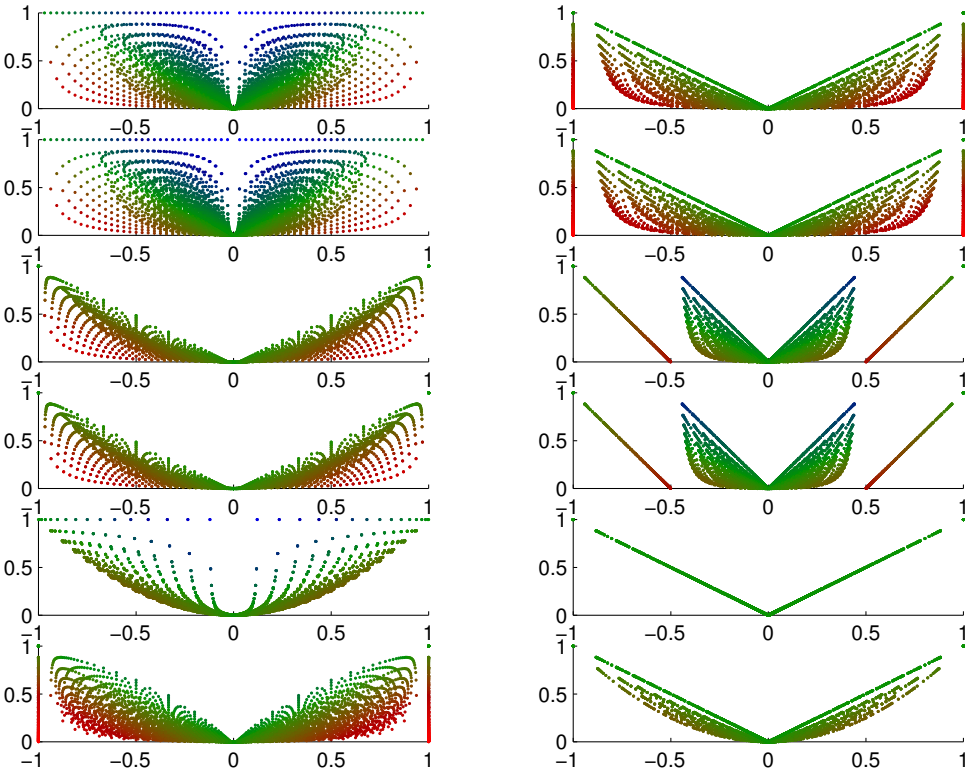

Figure 4. Triple-colour scatter-plots of the 12 selected confirmation measures against χ^2_{01} (left-hand column: measures $D(H,E)$, $M(H,E)$, $S(H,E)$, $N(H,E)$, $C(H,E)$, $F(H,E)$; right-hand column: measures $Z(H,E)$, $A(H,E)$, $c_1(H,E)$, $c_2(H,E)$, $c_3(H,E)$, $c_4(H,E)$)

- the levels of concordance are also generally reflected by the values of the s coefficient, although the ordering of the measures according to s is not exactly the same as that produced by r (e.g. measures $S(H,E)$ and $c_4(H,E)$).

As far as more detailed relations between the measures $c(H,E)$ and χ^2_{01}, as visualized by the scatter-plots of $c(H,E)$-versus-χ^2_{01}, are concerned, the following remarks can be made (see Figure 4):

- The $c(H,E)$-versus-χ^2_{01} scatter-plots are symmetric with respect to the vertical axis.

  This results from the fact that each of the selected measures $c(H,E)$ is symmetric with respect to 0 and in the exhaustive and non-redundant set of contingency tables there exist two different contingency tables producing opposite values of $c(H,E)$ and the same value of χ^2_{01}. This also causes the histograms of the measures to be symmetric with respect to 0 (see Figure 5).

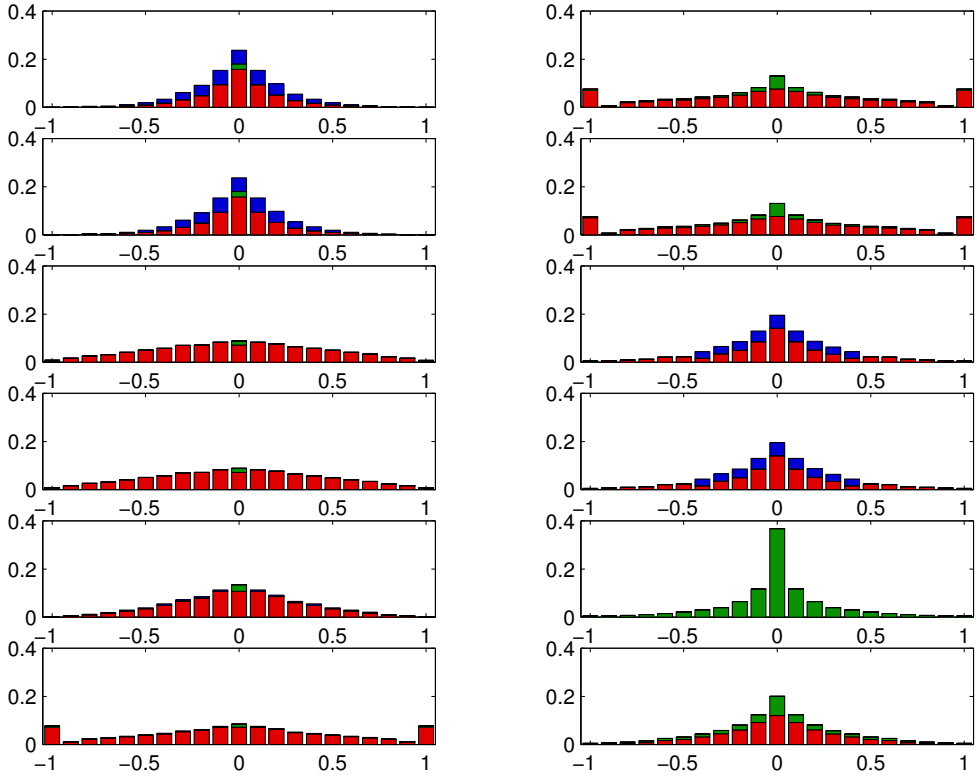

Figure 5. Triple-colour histograms of the 12 selected confirmation measures $c(H, E)$ in relation to χ^2_{01} (left-hand column: measures $D(H, E)$, $M(H, E)$, $S(H, E)$, $N(H, E)$, $C(H, E)$, $F(H, E)$; right-hand column: measures $Z(H, E)$, $A(H, E)$, $c_1(H, E)$, $c_2(H, E)$, $c_3(H, E)$, $c_4(H, E)$)

- The following pairs of confirmation measures generate the same $c(H, E)$-versus-χ^2_{01} scatter-plots: $D(H, E)$ and $M(H, E)$, $S(H, E)$ and $N(H, E)$, $Z(H, E)$ and $A(H, E)$, $c_1(H, E)$ and $c_2(H, E)$.

  This results from the fact that for each of those pairs, in the exhaustive and non-redundant set of contingency tables there exist contingency tables producing the same values of the respective measures and the same value of χ^2_{01}. This also causes the histograms of the respective pairs of measures to be the same (see Figure 5).

- The $c(H, E)$-versus-χ^2_{01} scatter-plot of the confirmation measures $Z(H, E)$ and $A(H, E)$ are special cases of the scatter-plots of $c_1(H, E)$ and $c_2(H, E)$ generated for $\alpha \to 1.0$.

  This results from the fact that for $\alpha \to 1.0$, $c_1(H, E) \to Z(H, E)$ and $c_2(H, E) \to A(H, E)$, thus the scatter-plots of all four measures converge in this case.

Let us observe that for $\alpha \to 0.0$ the $c(H,E)$-versus-χ^2_{01} scatter-plots of the confirmation measures $c_1(H,E)$ and $c_2(H,E)$ tend to resemble in some respects the scatter-plot of $c_3(H,E)$-versus-χ^2_{01}, which of course does not imply that $c_1(H,E) \to c_3(H,E)$ or $c_2(H,E) \to c_3(H,E)$. Generally, the scatter-plots of $c_1(H,E)$ and $c_2(H,E)$ consist of three segments that correspond to intervals $[-1,-\alpha)$, $(-\alpha,+\alpha)$ and $(+\alpha,+1]$. For $\alpha \to 0.0$ the interval $(-\alpha,+\alpha)$ degenerates to a single value, so the respective segment of the scatter-plot degenerates to a vertical line. At the same time, the remaining segments of the scatter-plots of $c_1(H,E)$ and $c_2(H,E)$ converge to the scatter-plot of $c_3(H,E)$.

Moreover, relations between the 12 selected confirmation measures and χ^2_{01} can have the following, risk-related interpretations (see Figure 5):

- every confirmation measure is risk-neutral whenever it obtains values equal to 0 (this concerns also measures other than the 12 selected confirmation measures being analysed in this paper),
- the confirmation measures $D(H,E)$, $M(H,E)$, $C(H,E)$ and $F(H,E)$ are predominantly risk-prone,
- the confirmation measures $S(H,E)$ and $N(H,E)$ are risk-prone in $(-1,0) \cup (0,+1)$ and risk-neutral for -1 and $+1$,
- the confirmation measures $Z(H,E)$ and $A(H,E)$ are never risk-averse,
- the confirmation measures $c_1(H,E)$ and $c_2(H,E)$ are predominantly risk-prone in $(-\alpha,+\alpha)$, risk-prone in $(-1,-\alpha) \cup (+\alpha,+1)$ and risk-neutral for -1 and $+1$,
- the confirmation measure $c_3(H,E)$ is always risk-neutral,
- the confirmation measure $c_4(H,E)$ is never risk-averse.

It it important to notice that the measures $c_1(H,E)$ and $c_2(H,E)$ depend on the value of the α parameter, i.e. the free parameter that is used to define these measures (the β parameter is, on the other hand, constrained, as $\beta = 1 - \alpha$). This necessarily influences the relations between these measures and the χ^2_{01} coefficient.

In particular, the parameter α influences the risk-related properties of $c_1(H,E)$ and $c_2(H,E)$. Firstly, the value set of measures $c_1(H,E)$ and $c_2(H,E)$ is $[-1,-\alpha) \cup (-\alpha,+\alpha) \cup (+\alpha,+1]$, which means that these measures are never equal to either $-\alpha$ or $+\alpha$ (see [31]). Now, the size of the intervals in which the measures are risk-prone, i.e. $(-1,-\alpha)$ and $(+\alpha,+1)$, is directly controlled by α. However, it should be stressed that the values of $c_1(H,E)$ and $c_2(H,E)$ belong to these intervals only when $c = 0$ or when $b = 0$, respectively, which strongly limits the number of such cases. In sets of all contingency tables satisfying $a+b+c+d = n$, as used in the experiments, this number is given by $(n+1)(n+2)/2$, and thus grows slower than the total number of tables t – see Table 4. This means that the situation will hardly ever happen for large n, e.g. for $n = 512$, $(n+1)(n+2)/2 = 131,841 \ll t = 22,632,705$.
Correspondingly, the profiles of the measures $c_1(H,E)$ and $c_2(H,E)$ are mostly determined by their behaviour in the interval $(-\alpha,+\alpha)$, which contains the majority of

Table 4. The percentages of the contingency tables of n observations characterized by $c = 0$, with n changing according to $n = 2^i$, where $i = 1..10$

n	t	$\lvert(c = 0)\rvert$	$\frac{100 \cdot \lvert(c=0)\rvert}{t}$ [%]
2	10	6	60.00
4	35	15	42.86
8	165	45	27.27
16	969	153	15.79
32	6,545	561	8.57
64	47,905	2,145	4.48
128	366,145	8,385	2.29
256	2,862,209	33,153	1.16
512	22,632,705	131,841	0.58
1,024	180,007,425	525,825	0.29

points (for $n > 3$). In result, e.g. for $\alpha = 0.50$, about 64% of all the points of the $c(H, E)$-versus-χ^2_{01} scatter-plots are situated below the $|c(H, E)| = \chi^2_{01}$ line (so the measures are predominantly 'risk-prone' for this α). Moreover, as it can be observed in Figure 6, $c_1(H, E)$ and $c_2(H, E)$ are predominantly risk-prone for $\alpha \gtrapprox 0.3$ and predominantly risk-averse for $\alpha \lessapprox 0.3$. This means that the α parameter can be directly used to control the profiles of these measures.

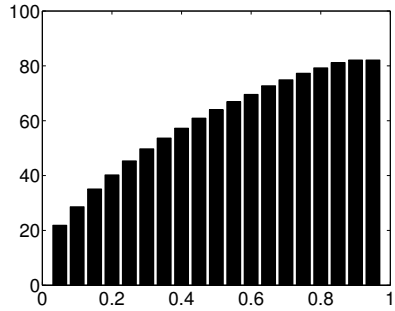

Figure 6. The percentages of points (the vertical axis) of the $c(H, E)$-versus-χ^2_{01} scatter-plots situated below the $|c(H, E)| = \chi^2_{01}$ line for $c_1(H, E)$ and $c_2(H, E)$ for α ranging from 0 to 1 (the horizontal axis)

Another interesting observation concerns the confirmation measure $c_3(H, E)$. This measure happens to satisfy $|c_3(H, E)| = \chi^2_{01}$, which implies $|c_3(H, E)| \cdot n = \chi^2_{01} \cdot n$, so $|\chi^2_{01}| \cdot n = \chi^2_0$. The assumed $n > 0$ means that the distribution of $|c_3(H, E)| \cdot n$ is known, as being equal to the distribution of χ^2_0, which is approximately χ^2-distributed (with $df = 1$ degree of freedom), so it is possible to test for the statistical significance of both $|c_3(H, E)| \cdot n$ and $|c_3(H, E)|$. The testing procedure is thus analogous to that of testing for the independency of two discrete-valued variables.

5. Concordance-driven exploitation of confirmation measures

As far as the broadly-taken concordance with the χ^2_{01} coefficient is concerned, confirmation measures (including measures other than the 12 selected measures being analysed in this paper), as analysed in the context of possibly erroneous observational data, may be generally categorized into two groups:

- autonomous,
- non-autonomous.

Measures referred to as autonomous are measures with distributions that are known to be identical (or at least similar) to the distribution of χ^2_{01}. A general result of the study indicates that this identity (or similarity) makes it possible to explicitly conclude about the significance of the dependency between the evidence and the hypothesis, which may directly translate to the usefulness of the measures (as their values have straightforward interpretations). It also allows to draw immediate conclusions as far as the risk-related profile of measures is concerned.

On the other hand, measures referred to as non-autonomous are measures of unknown distributions. However, another general result of the study indicates that 'parallel' analysis of such measure and the χ^2_{01} coefficient may also be useful, as computing the value of χ^2_{01} for the same data for which the measure was computed and testing the significance of this coefficient may also provide ample characterization of the measure. Additionally, a further comparison of the measure with the χ^2_{01} coefficient (e.g. in the form of a scatter-plot) may effectively reveal the risk-related profile of the measure. This, in turn, allows to identify its potentially non-neutral (in terms of risk) values. In result, the confirmation measure may be successfully handled, despite the fact that its actual distribution may remain unknown.

The above-mentioned idea of 'parallel' analysis of a confirmation measure and the χ^2_{01} coefficient may also be further exploited to deliver a new definition of whether a hypothesis is confirmed/disconfirmed by the evidence or not. Given a significance level α_0, a hypothesis H is:

- confirmed by the evidence E, when $c(H,E) > 0$ and χ^2_{01} is significant at α_0,
- neither confirmed nor disconfirmed by the evidence E, when χ^2_{01} is not significant at α_0 (regardless of the value of $c(H,E)$),
- disconfirmed by the evidence E, when $c(H,E) < 0$ and χ^2_{01} is significant at α_0.

On the other hand, the definition of a given confirmation measure may be modified directly to influence its properties. This modification may increase the concordance of the measure (as quantified e.g. by the Pearson correlation coefficient). In result, the measure will be more autonomous, which means that its values will have better interpretations in terms of risk. In particular, possible operations performed on single confirmation measures include:

- applying powers[8],

[8] Applying powers is used in the very definition of $Z(H,E)$ [6].

- multiplying by χ^2_{01} or by $1-p$, where $p = P(\chi^2 \geq |\chi^2_{01}| \cdot n)$.

As an example, let us consider $S^3(H,E) = (S(H,E))^3$ and $S_{\chi^2_{01}}(H,E) = S(H,E) \cdot \chi^2_{01}$ computed for $n = 128$ (see Table 5). Notice that the dependencies between the modified measures and the χ^2_{01} coefficient are generally better (i.e. higher as far as the r and s correlations are considered) than those between the original ones and χ^2_{01}.

Table 5. Correlation r and s coefficients of the original and modified confirmation measures $c(H,E)$

$c(H,E)$	$r(\lvert c(H,E)\rvert, \chi^2_{01})$	$s\lvert c(H,E)\rvert, \chi^2_{01})$
$S(H,E)$	0.912	0.960
$S^3(H,E)$	0.928	0.960
$S_{\chi^2_{01}}(H,E)$	0.975	0.995

At the same time, possible operations performed on multiple confirmation measures include:

- additive aggregation of confirmation measures,
- multiplicative aggregation of confirmation measures.

As an example, let us consider $c_1(H,E)$ and $c_3(H,E)$ computed for $n = 128$ (see Table 6). Measure $c_1(H,E)$ has been created as a basically additive, while measure $c_3(H,E)$ as a basically multiplicative aggregation of $Z(H,E)$ and $A(H,E)$. Also here the dependencies between the resulting measures and the χ^2_{01} coefficient are better, i.e. higher as far as the r and s correlations are considered (although the differences are rather low in the case of $c_1(H,E)$) than those between the component ones and χ^2_{01}.

Table 6. Correlation r and s coefficients of the component and modified confirmation measures $c(H,E)$

$c(H,E)$	$r(\lvert c(H,E)\rvert, \chi^2_{01})$	$s(\lvert c(H,E)\rvert, \chi^2_{01})$
$Z(H,E)$	0.694	0.782
$A(H,E)$	0.694	0.782
$c_1(H,E)$	0.697	0.783
$c_3(H,E)$	1.000	1.000

6. Conclusions

The paper considers a group of interestingness measures, called (Bayesian) confirmation measures. Given data patterns, such as decision rules, these measures express

the degree to which the rules' premises confirm their conclusions. In result, they allow to evaluate and rank the rules according to their interestingness, thus identifying the best ones (as well as the worst ones). Such an evaluation is a valid and useful step in approaches like data mining, in which different types of procedures, like model regularization and preference modelling meet.

The choice of an interestingness measure for a particular problem is a non-trivial task that should be preceded by thorough analyses of the measure's properties. Being aware of the measure's behaviour in particular situations (expressed by its properties) helps making such choice properly. Thus, the analysis of confirmation measures with respect to their properties is an active research area.

However, virtually all the studies of properties were confined to environments that had been explicitly or implicitly assumed to be free from observational errors. In real-life situations, however, the existence of such errors must be taken into account (as potentially false conclusions might be drawn otherwise) and this is what is assumed in this paper. The assumption about the environments being potentially error-prone requires a new rule evaluation approach, which amounts to determining whether the number of observations for or against a given rule is large enough in relation to all other observations, so that accidental conclusions could be eliminated. This goal, generally achievable with different tools, is in this paper particularly accomplished with the two-dimensional χ^2 test, commonly used to test for the dependency of two discrete-valued variables.

The actual amount of how much a confirmation measure 'agrees' with the level of dependency between the evidence and the hypothesis is quantified with the Pearson correlation coefficient between the measure and an introduced χ^2_{01} coefficient. High agreement is termed as concordance, and confirmation measures of high agreement with the coefficient are referred to as concordant ones. Thus, for each particular measure, it is the concordance that carries the actual answer to the question asked in the title of the paper.

The relations between a given confirmation measure and χ^2_{01} are additionally quantified by the Spearman rank correlation coefficient and illustrated by scatter-plots and specialized, triple-colour histograms. They are also interpreted in terms of risk. Following this interpretation, a confirmation measure that under-estimates the dependency between the evidence and the hypothesis is referred to as risk-averse, while a confirmation measure that over-estimates this dependency is referred to as risk-prone.

To quantify the concordance of confirmation measures in practical cases, a set of 12 particular measures has been selected and experimentally evaluated on pre-defined data. These data consist of all possible contingency tables having the same number of observations, constituting an exhaustive and non-redundant dataset of tables. Being exhaustive and non-redundant, this purposefully generated set of data ensures fairly uniform covering of the underlying space and natural interpretations of the results in probabilistic terms, while being deterministic, it ensures full repeatability of the results.

For the exhaustive and non-redundant dataset, the paper presents various forms of quantifications and illustrations of the relation between the measures and χ^2_{01}. After determining the actual relations between a given confirmation measure and the χ^2_{01}

coefficient, the measures have been categorized into more or less concordant, and the less concordant measures extensively characterized in terms of risk. Additionally, possible transformations of measures aimed at improving their concordance have been presented and discussed.

We postulate that all of the analyses introduced and described in this paper are general enough to be applied to basically all newly defined confirmation measures, which potentially could improve the general comprehension of the measures and, as such, may positively influence the way in which they will be defined in the future.

Potential further lines of investigations could include: considering alternative forms of testing for the dependency in data as well as alternative forms of expressing the concordance, designing a batch of real-life data experiments aimed at verifying the practical usability of concordant measures and extending the concordance analyses beyond the Bayesian confirmation measures.

References

[1] Bell, C., Mutual information and maximal correlation as measure dependence, *The Annals of Mathematical Statistics*, **33**, 1962, 587–595.

[2] Bjerve, S., Doksum, K., Correlation curves: Measures of association as functions of covariate value, *The Annals of Statistic*, **21**, 1993, 890–902.

[3] Brzezińska, I., Greco, S., Słowiński, R., Mining pareto-optimal rules with respect to support and anti-support, *Engineering Applications of Artificial Intelligence*, **20**, 5, 2007, 587–600.

[4] Carnap, R., *Logical Foundations of Probability, 2nd ed.*, University of Chicago Press, 1962.

[5] Christensen, D., Measuring confirmation, *Journal of Philosophy*, **96**, 1999, 437–461.

[6] Crupi, V., Tentori, K., Gonzalez, M., On bayesian measures of evidential support: Theoretical and empirical issues, *Philosophy of Science*, **74**, 2007, 229–252.

[7] Earman, J., *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory*, MIT Press, Cambridge, MA, 1992.

[8] Eells, E., Fitelson, B., Symmetries and asymmetries in evidential support, *Philosophical Studies*, **107**, 2, 2002, 129–142.

[9] Everitt, B., *The Analysis of Contingency Tables*, Chapman & Hall, 1992.

[10] Fitelson, B., The plurality of bayesian measures of confirmation and the problem of measure sensitivity, *Philosophy of Science*, **66**, 1999, 362–378.

[11] Fitelson, B., *Studies in Bayesian Confirmation Theory*, Ph.D. thesis, University of Wisconsin, Madison, 2001.

[12] Geng, L., Hamilton, H., Interestingness measures for data mining: A survey, *ACM Computing Surveys*, **38**, 3, 2006.

[13] Glass, D.H., Confirmation measures of association rule interestingness, *Knowlegde Based Systems*, **44**, 2013, 65–77.

[14] Greco, S., Pawlak, Z., Słowiński, R., Can bayesian confirmation measures be useful for rough set decision rules?, *Engineering Applications of Artifficial Intelligence*, **17**, 2004, 345–361.

[15] Greco, S., Słowiński, R., Szczęch, I., Properties of rule interestingness measures and alternative approaches to normalization of measures, *Information Sciences*, **216**, 2012, 1–16.

[16] Greco, S., Słowiński, R., Szczęch, I., Measures of rule interestingness in various perspectives of confirmation, *Inf. Sci.*, **346-347**, 2016, 216–235, doi:10.1016/j.ins.2016.01.056.

[17] Hämäläinen, W., Statapriori: an efficient algorithm for searching statistically significant association rules, *Knowl. Inf. Syst.*, **23**, 3, 2010, 373–399.

[18] Hämäläinen, W., Kingfisher: an efficient algorithm for searching for both positive and negative dependency rules with statistical significance measures, *Knowl. Inf. Syst.*, **32**, 2, 2012, 383–414.

[19] Hastie, T., Tibshirani, R., Friedman, J., *Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, 2003.

[20] Hempel, C., Studies in the logic of confirmation (i), *Mind*, **54**, 1945, 1–26.

[21] Joyce, J., *The Foundations of Causal Decision Theory*, Cambridge University Press, 1999.

[22] Keeney, R., Raiffa, H., *Decisions With Multiple Objectives: Preferences and Value Tradeoffs*, John Wiley and Sons, Inc, 1976.

[23] Kemeny, J., Oppenheim, P., Degrees of factual support, *Philosophy of Science*, **19**, 1952, 307–324.

[24] Lenca, P., Meyer, P., Vaillant, B., Lallich, S., On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid, *European Journal of Operational Research*, **184**, 2, 2008, 610–626.

[25] Luce, R., Raiffa, H., *Games and Decisions: Introduction and Critical Survey*, John Wiley and Sons, Inc, 1957.

[26] Mortimer, H., *The Logic of Induction*, Paramus, Prentice Hall, 1988.

[27] Nozick, R., *Philosophical Explanations*, Clarendon Press, Oxford, UK, 1981.

[28] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufman, San Francisco, 1988.

[29] Rayner, J., Best, D., *A Contingency Table Approach to Nonparametric Testing*, Taylor & Francis Group, 2001.

[30] Słowiński, R., Szczęch, I., Urbanowicz, M., Greco, S., Mining association rules with respect to support and anti-support-experimental results, *Rough Sets and Intelligent Systems Paradigms, International Conference, RSEISP 2007, Warsaw, Poland, June 28-30, 2007, Proceedings*, 2007, 534–542.

[31] Susmaga, R., Szczęch, I., Statistical significance of bayesian confirmation measures, Technical report, RA-010/12, Poznań University of Technology, 2012.

[32] Susmaga, R., Szczęch, I., The property of χ^2_{01}-concordance for bayesian confirmation measures, *Lecture Notes in Artificial Intelligence*, **8234**, 2013, 226–236.

[33] Susmaga, R., Szczęch, I., Visualization support for the analysis of properties of interestingness measures, *Bulletin of the Polish Academy of Sciences. Technical Sciences*, **63**, 1, 2015, 315–327.

[34] Susmaga, R., Szczęch, I., Selected group-theoretic aspects of confirmation measure symmetries, *Inf. Sci.*, **346-347**, 2016, 424–441, doi:10.1016/j.ins.2016.01.041.

[35] Szczęch, I., Multicriteria attractiveness evaluation of decision and association rules, *Transactions on Rough Sets X, LNCS series*, **5656**, 2009, 197–274.

[36] Tentori, K., Crupi, V., Bonini, N., Osherson, D., Comparison of confirmation measures, *Cognition*, **103**, 2007, 107–119.

[37] Tick, J., Fodor, J., Fuzzy implications and inference processes, *Computing and Informatics*, **24**, 2005, 591–602.

[38] Venables, W., Ripley, B., *Modern Applied Statistics with S*, Springer-Verlag, 2002.

Received 29.09.2017, Accepted 30-01-2018