Visual-based Detection of Properties of Confirmation Measures

Robert Susmaga and Izabela Szczęch

Institute of Computing Science, Poznań Univesity of Technology,
Piotrowo 2, 60-965 Poznań, Poland

Abstract. The paper presents a visualization technique that facilitates and eases analyses of interestingness measures with respect to their properties. Detection of properties possessed by these measures is especially important when choosing a measure for KDD tasks. Our visual-based approach is a useful alternative to often laborious and time consuming theoretical studies, as it allows to promptly perceive properties of the visualized measures. Assuming a common, four-dimensional domain of the measures, a synthetic dataset consisting of all possible contingency tables with the same number of observations is generated. It is then visualized in 3D using a tetrahedron-based barycentric coordinate system. Additional scalar function – an interestingness measure – is rendered using colour. To demonstrate the capabilities of the proposed technique, we detect properties of a particular group of measures, known as confirmation measures.

Keywords: Visualization, interestingness measures, confirmation measures, properties of measures

1 Introduction

Within data mining tasks, a valid phase concentrates on the evaluation of induced patterns, in the form of e.g., "*if* premise, *then* conclusion" rules. The number of rules induced even from relatively small datasets can be overwhelming for users/decision makers and needs to be limited, so that irrelevant or misleading patterns can be discarded. Such an evaluation is commonly done by means of interestingness measures, e.g., measures of support and confidence for association rules or confirmation measures for decision rules. The variety of measures of interest proposed in the literature makes it, however, difficult to choose a measure or a group of measures (in case of a multi-criteria evaluation) for a particular application.

To help users make that choice, properties of measures are intensively studied [4, 7, 9, 10]. Theoretical analyses of measures with respect to their properties are often time consuming and require laborious mathematical calculations. In this paper we adopt (similarly to [11]) a visualization technique that aids the process of analysing measure properties. It allows to conduct a preliminary detection of properties satisfied by each visualized measure. In particular, it facilitates finding counterexamples, i.e., examples that discard particular properties. Our visual-based property detection eases the analysis not only for the measures already known in the literature, but also for newly developed ones (e.g., automatically generated).

In this paper, the visualization technique is adopted to detection of properties of a particular group of measures, called *confirmation measures*. These measures are designed to quantify the strength of confirmation that a rule's premise gives to its conclusion, and are thus a commonly used tool for the evaluation of decision rules [3, 6]. The particular properties that we will focus on in this paper are:

- property M, of monotonic dependency of the measure on the number of objects supporting or not the rule's premise or conclusion [6, 12],
- property of *maximality/minimality*, indicating the necessary and sufficient conditions under which measures should obtain their extreme values [5],
- property of hypothesis symmetry HS, stating that valuable confirmation measures should obtain the same value but of opposite sign for rules $E \rightarrow H$ and $E \rightarrow \neg H$ [1, 2, 5, 8]. Hypothesis symmetry is actually one of many symmetry properties considered in the literature, but it is the only one that all the authors agree to find truly desirable.

Let us observe that the chosen properties also reflect many other properties, such as the well known Piatetsky-Shapiro's properties [10]. For example, the Piatetsky-Shapiro's requirement for a measure to obtain value 0 when the premise and conclusion are statistically independent corresponds to a part of the definition of confirmation (see Section 2). Similarly, given a rule $E \rightarrow H$, the requirement for a measure to monotonically increase with $P(E, H)$ when $P(E)$ (or $P(H)$) remains the same coincides with property M; analogously, the requirement that a measure monotonically decreases with $P(E)$ (or $P(H)$) when $P(E, H)$ and $P(H)$ (or $P(E)$) remain the same is included in property M(see Section 3). One can thus conclude that the property of confirmation and property M extend the requirements known as the Piatetsky-Shapiro's properties.

The rest of the paper is organized as follows. A selection of popular confirmation measures is described in Section 2, followed by their commonly studied properties in Section 3. Next, Section 4 demonstrates the proposed visualization technique, which is then employed to visual-based analyses of the selected measures with respect to those properties in Section 5. Final remarks are collected in Section 6.

2 Confirmation measures

The confirmation measures, some of which are presented and analysed in this paper, constitute an important group among measures of interest. They evaluate rule patterns induced from a set of objects U and described by a set of attributes with respect to their relevance and utility [3, 5, 7, 8]. Formally, for a rule $E \rightarrow H$, an interestingness measure $c(H, E)$ is a confirmation measure when it satisfies the following conditions:

$$c(H,E)\begin{cases} >0 \text{ when } P(H|E)>P(H),\\ =0 \text{ when } P(H|E)=P(H),\\ <0 \text{ when } P(H|E)<P(H).\end{cases} \tag{1}$$

Confirmation is, thus, regarded as an increase in the probability of the conclusion H provided by the premise E (similarly for neutrality and disconfirmation).

In practical applications, where frequentionist's approach is often employed, the relation between E and H is quantified by four non-negative numbers:

- a: the number of objects in U for which both E and H hold,
- b: the number of objects in U for which $\neg E$ and H hold,
- c: the number of objects in U for which E and $\neg H$ hold,
- d: the number of objects in U for which $\neg E$ and $\neg H$ hold.

Naturally, a, b, c and d all sum up to n, being the cardinality of the dataset U. Working with a, b, c and d has two advantages, namely they can be used to estimate probabilities e.g., the conditional probability of the conclusion given the premise is $P(H|E) = P(H \cap E)/P(E) = a/(a+c)$, and, secondly, they can be used to calculate the value of confirmation measures, as each of them is a scalar function of a, b, c and d. From a long list of alternative and oridinally non-equivalent confirmation measures, let us concentrate on six exemplary ones, defined in Table 1 (the selection inspired by [5, 6]).

Table 1. Popular confirmation measures

$M(H,E) = P(E\|H) - P(E) = \frac{a}{a+b} - \frac{a+c}{n}$
$S(H,E) = P(H\|E) - P(H\|\neg E) = \frac{a}{a+c} - \frac{b}{b+d}$
$F(H,E) = \frac{P(E\|H) - P(E\|\neg H)}{P(E\|H) + P(E\|\neg H)} = \frac{ad-bc}{ad+bc+2ac}$
$FS(H,E) = \frac{1}{2}(F(H,E) + S(H,E))$
$\phi(H,E) = \frac{P(E,H) - P(E)P(H)}{\sqrt{P(E)P(H)P(\neg E)P(\neg H)}} = \frac{ad-bc}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$
$F\phi(H,E) = \frac{1}{2}(F(H,E) + \phi(H,E))$

The domains of the six selected measures range from -1 to $+1$, allowing us to translate them later on into a predefined colour map. Let us also note that the idea behind the formulation of measures $FS(H,E)$ and $F\phi(H,E)$ is that they benefit from properties of their constituent measures [5].

To better characterize the measures and to help choose suitable ones for the user, many properties have been proposed and compared in the literature [2, 4, 5, 7, 8]. The definitions of the most commonly required ones are presented in the next section, followed by visual-based analyses of measures from Table 1 with respect to these properties.

3 Properties of confirmation measures

Property of monotonicity M

Property of monotonicity M, introduced by Greco, Pawlak and Słowiński in [6], requires that a confirmation measure $c(H, E)$ is a function:

- non-decreasing with respect to a and d, and
- non-increasing with respect to b and c.

As a result, measures possessing M cannot decrease its value when a new positive example is introduced to the dataset (i.e., a increases). Similarly for objects not satisfying neither the rule's premise nor its conclusion (i.e., increase of d). At the same time, when new counterexamples are put into the dataset (i.e., increase of c), then a measure possessing property M cannot increase its value (analogously with b).

Property of maximality/minimality

The property of maximality/minimality has been introduced by Glass in [5] as a requirement stating necessary and sufficient conditions under which measures should obtain their extreme values. Formally, provided a confirmation measure $c(H, E)$ is defined, it enjoys property of maximality/minimality when:

- $c(H, E)$ is maximal if and only if $P(E, \neg H) = P(\neg E, H) = 0$, and
- $c(H, E)$ is minimal if and only if $P(E, H) = P(\neg E, \neg H) = 0$.

Equivalently, a measure satisfying the maximality/minimality requirement obtains its maximum iff $c = b = 0$, and obtains its minimum iff $a = d = 0$ [5]. Let us observe that the maximality/minimality property is closely related to properties Ex_1 [1] and weak Ex_1 [7, 8], as well as to properties L [1, 3] and weak L [7, 8].

Property of hypothesis symmetry

In the literature there is an intensive discussion about a whole group of symmetry properties, where each symmetry considers how the value of a confirmation measure $c(H, E)$ relates to its value obtained for the situation in which the rule's premise and/or conclusion is negated and/or when the premise and conclusion switch positions. The symmetry properties have been analysed, among others, by Eells and Fitelson [2], Crupi, Tentori and Gonzalez [1], Greco, Słowiński and Szczęch [8] as well as Glass [5]. They have different opinions as to which symmetries properties are desirable, but they all agree that a valuable confirmation measure should satisfy hypothesis symmetry HS stating that

$$c(H, E) = -c(\neg H, E).$$

This means that the value of a confirmation measure for rule $E \to H$ should be the same but of opposite sign as for a rule $E \to \neg H$. To present the intuition behind hypothesis symmetry, let us consider a rule *"if x is a square, then x is a rectangle"*. Obviously, the strength with which the premise (x is a square) confirms the conclusion (x is a rectangle) is the same as the strength with which the premise disconfirms the negated conclusion (x is not a rectangle). Thus, it is natural to expect that $c(H, E) = -c(\neg H, E)$.

4 The visualization technique

The visual-based detection of properties of confirmation measures employs, first of all, synthetic data that consist of an exhaustive and non-redundant set of contingency tables. Given a constant $n > 0$ (the total number of observations), it is generated as the set of all possible $\left[\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right]$ tables satisfying $a + b + c + d = n$. The set thus contains exactly one copy of each such table. We use $n = 128$ implying $t = 366145$ in all further visualizations.

Let us observe that using a synthetic dataset allows us to gain an insight into all areas that can be possibly occupied by the domain of a measure, including areas that could otherwise be omitted when using a real-life dataset. Thus, our approach reveals all possible behaviours and features of the considered measure. It also makes the results general, rather than application-specific.

The operational data set comprises t rows and 4 columns (representing a, b, c and d). Four columns correspond to four degrees of freedom, thus, visualization of such data would require four dimensions, however owing to the constraint $a + b + c + d = n$, the number of degrees of freedom is reduced to 3. Thus, it is possible to represent such data in three dimensions (3D) using tetrahedron-based barycentric coordinates. The tetrahedron, as used throughout the paper (and rendered in what will be referred to as the standard view), has its four vertices A, B, C and D coinciding with points of the following $[x, y, z]$ coordinates: A: $[1, 1, 1]$, B: $[-1, 1, -1]$, C: $[-1, -1, 1]$ and D: $[1, -1, -1]$.

The interpretation of the tetrahedron points is as follows: the vertex A corresponds to the (single) contingency table satisfying $a = n$ and $b = c = d = 0$, the edge AB corresponds to the (multiple) contingency tables satisfying $a + b = n$ and $c = d = 0$, etc. A standard view of a skeleton tetrahedron (only edges visible) is depicted in Figure 1.

For a more comprehensive visualization, the standard view will be accompanied by a rotated view, which depicts the DAB face of the tetrahedron (not visible in the standard view). The combination of these two views will be referred to as the 3D 2-view visualization of the tetrahedron.

Let us now observe that, since the points of the tetrahedron may be displayed in colour, the proposed visualization technique can also visualize a function $f(a, b, c, d)$ of the four arguments, e.g., any interestingness measure. It is assumed that the value set of this function is a real interval $[r, s]$, with $r < s$, so that its values may be rendered using a pre-defined colour map. The standard colour map used in the following visualizations is: from dark blue (corresponding

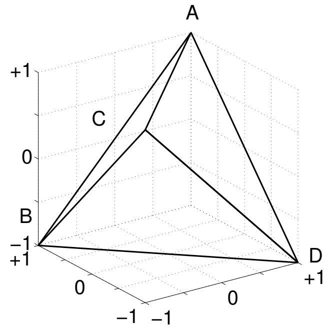

Fig. 1. A skeleton visualization of the tetrahedron

to r), through pale green, up to dark brown (corresponding to s). Non-numeric values, i.e., $+\infty$, NaN and $-\infty$, if generated by a particular function, may be rendered as colours not occurring in the map.

Notice that the 3D 2-view visualization of a 'solid' tetrahedron shows only extreme values of the arguments of the visualized function (external view). If areas located strictly inside the tetrahedron need to be additionally visualized, various variants of the visualization may be generated (internal views), however due to page limit, we shall not cover that topic in this paper.

5 Visual-based detection of properties of measures

The proposed visualization technique presents a coloured tetrahedron for each of the analysed confirmation measures. The values of the measures range from -1 to $+1$, thus a standard colour map with shades of dark blue (minimal values) through pale green (neutral values) to dark brown (maximal values) shall be used. The NaN values are rendered in magenta and depicted with special character ('$*$').

The following subsections explain how the visualization eases and speeds up the detection of particular properties of confirmation measures. It should be stressed, however, that essentially the same techniques may be applied to other measures, e.g. the group of measures that assess the performance of binary classification procedures (sensitivity, specificity, F-score, precision, accuracy, etc.). One of the fundamental differences between measures from these groups is that while the confirmation measures are bi-polar (i.e. their values range from -1 to $+1$), the performance measures are customarily uni-polar (i.e. their values range 0 to $+1$). Certainly, this does not prevent the performance measures from being visualized using the presented technique, however, the bi-polarity makes the confirmation measures and their analysis more difficult, and thus more interesting. In result, it is the confirmation measures that have been chosen to be discussed and visualized below.

5.1 Detection of property M

The definition of property of monotonicity M expects a confirmation measure $c(H,E)$ to be a function non-decreasing with respect to a and d, and non-increasing with respect to b and c. Such requirements directly translate to particular colour changes that are allowed at the visualized tetrahedrons. Precisely, the “non-decreasing with a and d” condition should be reflected in the visualization as colours changing towards dark brown (increase of confirmation) around vertices A and D and the “non-increasing with b and c” condition should be reflected in the visualization as colours changing towards dark blue (increase of disconfirmation) around vertices B and C. The proposed visualization technique allows thus to promptly perceive cases (counterexamples) that are contrary to the colour patterns required by property M. In fact, if there occurs any colour change that does not follow the expected gradient towards dark brown (dark blue) around vertices A and D (B and C), we can conclude that the visualized measure does not possess the property M. Let us stress, however, that potential counterexamples can also be hidden inside the tetrahedron, thus a thorough analysis of property M must take into account the inner parts of the shape as well.

Among the measures considered in this paper, clearly, measure $M(H,E)$, depicted in Figure 2 (row 1), does not satisfy property M, as in the visualization the colour changes from dark brown at vertex D to pale green at vertex A, violating the demands the of the non-decrease with a.

On the other hand, there are no observable counterexamples to property M in the external visualizations of other considered measures (see e.g., measure $S(H,E)$ in Figure 2 (row 2)) which, together with additional analysis of the inside of the tetrahedrons, determines the possession of the property by those measures.

5.2 Detection of maximality/minimality property

The maximality/minimality property requires that a confirmation measure obtains its maximal values if and only if $b = c = 0$. Additionally, minimal values are to be obtained if and only if $a = d = 0$. This translates into expectation that the dark brown (dark blue) colour must be found on the AD (BC) edge of the tetrahedron and cannot be found anywhere else. Let us observe that the AD (BC) edge contains all points for which $b = c = 0$ ($a = d = 0$), i.e., the points most distant from the vertices B and C (A and D). Similarly as with property M, it must be stressed that a thorough analysis with respect to maximality/minimality also requires an insight into the tetrahedron as potential counterexamples to this property may be located inside the shape.

Visual-based detection of maximality/minimality property among the measures in Table 1 reveals that measures $M(H,E)$ (see Figure 2 (row 1)) and $F(H,E)$ (see Figure 2 (row 3)) are the ones that do not satisfy this property. It is, among others, due to the fact that the points with maximal values of measure $M(H,E)$ are located only at vertex D of AD edge and thus do not cover the whole edge (i.e., there are too few of them), and the points with maximal values of measure $F(H, E)$ cover the whole ABD face (i.e., there are too many of them). The remaining measures do satisfy maximality/minimality (see e.g., measure $FS(H, E)$, presented in Figure 2 (row 4)). It is because the external views of the tetrahedra representing the corresponding measures do not depict counterexamples to this property and the additional insight into the shapes did not reveal any such cases either.

5.3 Detection of property of hypothesis symmetry

From definition, the hypothesis symmetry demands that a measure $c(H, E)$ obtains the same values, but of the opposite sign, for rules $E \rightarrow H$ and $E \rightarrow \neg H$, i.e., $c(H, E) = -c(\neg H, E)$. Let us assume that the first rule is characterized by a contingency table with numbers a, b, c and d, and the latter by a table with numbers a', b', c' and d'. Then, the conditions for hypothesis symmetry translate into: $c(H, E) = f(a, b, c, d) = -c(\neg H, E) = -f(a', b', c', d') = -f(c, d, a, b)$, reflecting the exchange of columns in the contingency tables ($a = c'$, $b = d'$, $c = a'$ and $d = b'$). In the context of our visualization technique, hypothesis symmetry detection boils down to the determination if the two presented views of a considered measure have the same gradient profile (i.e., if the left view is just like the right one, provided the colour map is reversed). If the 'recoloured' views are not the same, then the visualized measure does not possess the hypothesis symmetry. Again, a thorough analysis of HS requires an insight into the tetrahedron, as potential counterexamples can also lie inside the shape.

Figure 2 (row 1), shows counterexamples to hypothesis symmetry of measure $M(H, E)$, since e.g., the BCD face has a gradient profile that is characterized by straight lines running parallel to edge BD, while the DAB face has a profile that is characterized by curved lines coinciding with vertex D. Simple change of colours will, thus, not result in unifying faces BCD and ABD, which implies that measure $M(H, E)$ does not satisfy the hypothesis symmetry. On the other hand, there are no observable counterexamples to this property in other considered measures (see e.g., measure $\phi(H, E)$ in Figure 2 (row 5) and measure $F\phi(H, E)$ in Figure 2 (row 6)).

6 Conclusions

The choice of a particular interestingness measure for the evaluation of rules induced from datasets is often a difficult task. Thus, determination of properties possessed by measures became an active research area. This paper presents a visual technique designed to support and ease the detection of properties of measures. Such visual-based approach may be advantageous, especially when time constraints impede conducting in-depth, theoretical analyses of large numbers of such measures (e.g., generated in an automatic way). Our proposition starts with constructing a synthetic, exhaustive and non-redundant set of contingency tables, which are commonly used to calculate the values of measures.

Using such dataset, a 3-dimensional tetrahedron is built. The position of points in the shape translates to corresponding contingency tables and the colour of the points represents values of the visualized measure.

The proposed visualization allows us to promptly detect distinct properties of the measures and compare them, increasing the general comprehension of the measures and helping the users choose one for their particular application. For illustrative purposes, we have conducted and described a visual-based analysis of six popular confirmation measures with respect to three chosen properties. Clearly, the analyses can be generalized to a wider range of measures or properties.

Acknowledgment. The work was supported by local grant 09/91/DSPB/0543.

References

1. Crupi, V., Tentori, K., Gonzalez, M.: On bayesian measures of evidential support: Theoretical and empirical issues. Philosophy of Science 74, 229–252 (2007)
2. Eells, E., Fitelson, B.: Symmetries and asymmetries in evidential support. Philosophical Studies 107(2), 129–142 (2002)
3. Fitelson, B.: The plurality of bayesian measures of confirmation and the problem of measure sensitivity. Philosophy of Science 66, 362–378 (1999)
4. Geng, L., Hamilton, H.: Interestingness measures for data mining: A survey. ACM Computing Surveys 38(3) (2006)
5. Glass, D.H.: Confirmation measures of association rule interestingness. Knowlegde Based Systems 44, 65–77 (2013)
6. Greco, S., Pawlak, Z., Słowiński, R.: Can bayesian confirmation measures be useful for rough set decision rules? Eng. Appl. of Artifficial Intelligence 17, 345–361 (2004)
7. Greco, S., Słowiński, R., Szczęch, I.: Properties of rule interestingness measures and alternative approaches to normalization of measures. Information Sciences 216, 1–16 (2012)
8. Greco, S., Słowiński, R., Szczęch, I.: Finding meaningful bayesian confirmation measures. Fundamenta Informaticae 127(1-4), 161–176 (2013)
9. Hébert, C., Crémilleux, B.: A unified view of objective interestingness measures. In: Perner, P. (ed.) MLDM. Lecture Notes in Computer Science, vol. 4571, pp. 533–547. Springer (2007)
10. Piatetsky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Knowledge Discovery in Databases, pp. 229–248. AAAI/MIT Press (1991)
11. Susmaga, R., Szczęch, I.: Visualization of interestingness measures. In: Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics. pp. 95–99 (2013)
12. Szczęch, I.: Multicriteria attractiveness evaluation of decision and association rules. Transactions on Rough Sets X, LNCS series 5656, 197–274 (2009)

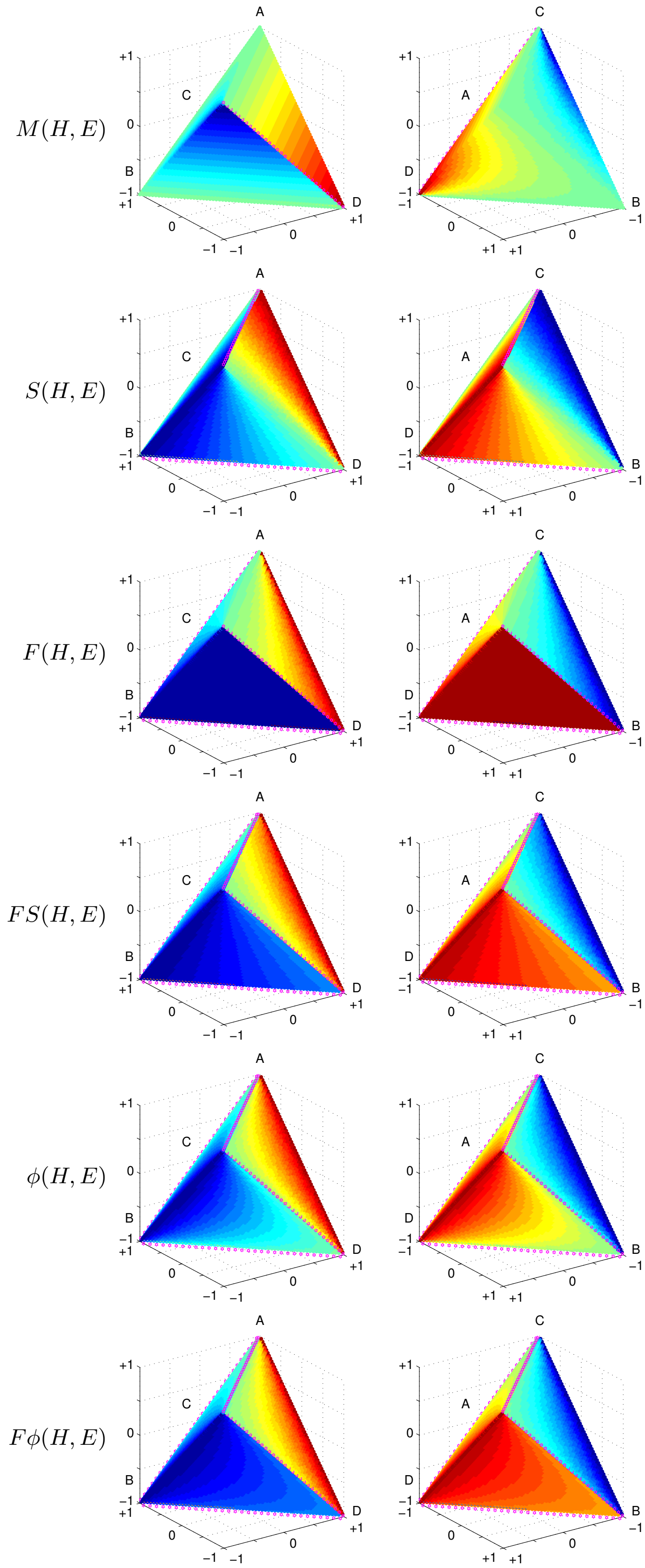

Fig. 2. A 3D 2-view visualization of the selected confirmation measures.