Visualization of Interestingness Measures

Robert Susmaga, Izabela Szczęch

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

Abstract

The paper presents visualization techniques for interestingness measures, which provide useful insights into different domain areas of the visualized measure and thus effectively assist measure comprehension and their selection for KDD methods. Assuming a common, 4-dimensional domain form of the measures, the system generates a synthetic set of contingency tables and visualizes them in three dimensions using a tetrahedron-based barycentric coordinate system. At the same time, an additional, scalar function of the data (referred to as the operational function, e.g. any interestingness measure) is rendered using colour. Throughout the paper a particular group of interestingness measures, known as confirmation measures, is used to demonstrate various capabilities of the visualization techniques, which range from the determination of specific values (extremes, zeros, etc.) of a single measure, to the localization of pre-defined regions of interest, e.g. such domain areas for which two/or more measures do not differ at all or differ the most.

Keywords: Visualization, interestingness measures, confirmation measures, barycentric coordinates

1. Introduction

Rapid progress in data mining and knowledge discovery techniques has increased over the recent years our ability to extract answers from data. Apparently, it influences intense work on data visualization to improve our abilities to present the information in meaningful ways. Data visualization can provide graphical displays for thorough data comprehension, it is thus natural that the development of KDD tools is accompanied by the development of various visualization techniques.

The paper presents visual techniques that support the analysis of interestingness measures commonly used to evaluate *if-then* rule mined from data (Fayyad et al., 2002). The rule induction process usually requires an evaluation step to limit the number of rules presented to the user and quantitative measures of interest are often used for such filtration. It is not easy, though, to choose an appropriate measure for a particular application. To help to do so, our techniques visualize the values obtained by a measure for a synthetic data set consisting of an exhaustive and non-redundant set of contingency tables. This way we gain a valuable insight into all areas of the domain that the visualized measure can possibly occupy.

The analyses facilitated by our MATLAB-based implementation of the techniques range from the determination of measure's extremes or the areas for which its value is undefined, to the visualization of the areas of the data set for which two or more measures differ the most. One could then e.g. decide to work with a couple of measures that react to different (types of) objects in the data set, or could choose to use measures that are not ordinally equivalent.

Besides enriching our theoretical knowledge on the features and the behaviour of the visualized measures, conclusions drawn on the basis of our visualization tool are also very practical, as they guide the user towards an interestingness measure (or measures) that best reflects his/her expectations. Moreover, our tool eases defining new measures and facilitates the analysis of newly developed ones (e.g. automatically generated).

In this paper, the exemplary application of our visualization techniques is presented for a particular group of interestingness measures called *confirmation measures*, designed for the evaluation of decision rules, in the form of "*if E, then H*", with E referring to an existing piece of evidence, and H referring to a hypothesised piece of evidence.

The confirmation measures are characterised by the fact that they obtain:

- values > 0 when the premise E of a rule confirms its conclusion H,
- values $= 0$ when the rule's premise E and conclusion H are neutral to each other,
- values < 0 when the premise E of a rule disconfirms its conclusion H.

In the context of a particular data set, the relation between E and H may be quantified with a 2×2 contingency table of non-negative frequencies a, b, c and d (see Table 1), where:

- a counts objects satisfying both the premise and the conclusion,
- b counts objects satisfying the premise but not the conclusion,
- c counts objects satisfying the conclusion but not the premise,
- d counts objects satisfying neither the premise nor the conclusion.

Let us observe that a, b, c and d can be used to estimate probabilities: e.g. the probability of the premise is expressed as $P(E) = (a + c)/n$, the conditional probability of the conclusion given the premise is $P(H|E) = P(H \cap E)/P(E) = a/(a + c)$, etc. Definitions of 6 popular confirmation measures (in terms of a, b, c and d) are presented in Table 2.

Table 2: Popular confirmation measures

$D(H,E) = P(H\|E) - P(H) = \frac{a}{a+c} - \frac{a+b}{n}$	(Eells, 1982)
$M(H,E) = P(E\|H) - P(E) = \frac{a}{a+b} - \frac{a+c}{n}$	(Mortimer, 1988)
$S(H,E) = P(H\|E) - P(H\|\neg E) = \frac{a}{a+c} - \frac{b}{b+d}$	(Christensen, 1999)
$N(H,E) = P(E\|H) - P(E\|\neg H) = \frac{a}{a+b} - \frac{c}{c+d}$	(Nozick, 1981)
$C(H,E) = P(E \wedge H) - P(E)P(H) = \frac{a}{n} - \frac{(a+c)(a+b)}{n^2}$	(Carnap, 1962)
$F(H,E) = \frac{P(E\|H) - P(E\|\neg H)}{P(E\|H) + P(E\|\neg H)} = \frac{ad-bc}{ad+bc+2ac}$	(Kemeny and Oppenheim, 1952)

Table 1: A exemplary contingency table of the rule's premise E and conclusion H

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

The rest of the paper is organized as follows. Section 2. demonstrates the proposed visualization techniques. Section 3. presents the application of the visualization techniques to popular confirmation measures defined in this Introduction. Exemplary conclusions drawn from the visualization-based analyses are also described. Final remarks and conclusions are contained in Section 4.

2. Visualization techniques

For the purpose of our visualization, a synthetic data set consisting of an exhaustive and non-redundant set of contingency tables has been prepared. 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. The total number of contingency tables t in the set is given by $t = (n+1)(n+2)(n+3)/6$. We use $n = 64$ (and thus $t = 47905$) in all further computations and visualizations. The resulting data set comprises t rows and 4 columns: a, b, c and d. Because, in general, four independent columns correspond to four degrees of freedom, visualization of such data in the form of a scatter-plot would formally require four dimensions. Owing to the condition $a+b+c+d = n$, however, the number of degrees of freedom is reduced to three, so it is possible to visualize such data in three dimensions (3D) using tetrahedron-based barycentric coordinates (Warren, 2003).

The 3D view of the tetrahedron, as used throughout the paper (and 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 combination of viewing angles (azimuth, elevation) in the standard view is $(-35\,°, 22\,°)$. It is accompanied by a rotated view, viewing angles $(145\,°, 22\,°)$, which depicts the DAB face of the tetrahedron (not visible in the standard view). The combination of these views will be collectively referred to as the 3D 2-view visualization of the tetrahedron. 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 $c = d = 0$, the face ABC corresponds to the (multiple) contingency tables satisfying $a+b+c = n$ and $d = 0$, etc.

Because the individual points of the tetrahedron may be displayed in colour, it is possible to visualize a function $f(a,b,c,d)$ of the four arguments, further referred to as the operational function (e.g. any interestingness measure). It is additionally 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[1] used in the following visualizations is: from dark blue (corresponding 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 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 have to be additionally visualized, various variants of the visualization may be generated (internal views).

Summarizing, the capabilities of the visualization techniques include:

- regular views of any operational function,
- specialized views of a region of interest, i.e. only points satisfying pre-defined conditions, e.g. $f(a,b,c,d) = 0$, of any operational function,

[1] Owing to the printing restrictions, the standard colour map had to be substituted with a grey colour map, with black and white corresponding to -1 to $+1$, respectively. Additionally, to increase the clarity of the presentation, some values have been depicted with special characters ('$+$', '$*$', etc.).

- specialized views of any number of operational functions
  - differences between two operational functions,
  - variances/means of a number of operational functions.

3. Application of the visualization techniques

3.1. Regular views of confirmation measures

Taking particular confirmation measures as operational functions, the regular views of the measures may be used to practically compare their general configurations of values and gradient profiles. Consider exemplary external visualizations of measures $S(H,E)$, $C(H,E)$ and $F(H,E)$, as presented[2] in Figures 1, 2 and 3. Such visualizations potentially allow to instantly notice fundamental differences, e.g. between their gradient profiles. Observe that in all their faces measures $S(H,E)$ and $C(H,E)$ manifest 'radial' and 'concentric' gradients, respectively, while measure $F(H,E)$ is characterized by constant values (and thus no gradient) in two faces (ABD and BCD) and a 'radial' gradient in the other two. Such visual analyses allow to tentatively conclude about the ordinal equivalence of the visualized measures, an especially important issue in evaluating rules with multiple measures. In the case of $S(H,E)$, $C(H,E)$ and $F(H,E)$ the different gradient profiles in the external areas of the corresponding tetrahedrons constitute conclusive counterexamples to the ordinal equivalence of those measures. In general, however, this kind of equivalence analysis may require an insight into the interior of the tetrahedron.

3.2. Specialized views of regions of interest

In their analyses of the confirmation measures, users may be interested in discovering regions for which the considered measures satisfy some pre-defined conditions, e.g. $c(H,E) = 0$ (the neutral value) or $c(H,E) = +1$ (the maximal value). Supporting the user with such specialized views is important for at least two reasons: it allows to test for the existence of such regions and to identify the localizations of these regions within the tetrahedron, translating them uniquely to particular values of a, b, c and d.

Figure 4 depicts[3] regions for which $|C(H,E)| = 0.5$. Notice the full symmetry of these regions (coinciding with edges BC and AD, while avoiding the other edges).

Other exemplary regions of interest are presented in Figure 5. It depicts both extreme (-1 and $+1$) and non-numeric values (NaN) of measure $N(H,E)$. Notice that the non-numeric values exist in two disjoint localizations, i.e. in edges AB and CD in the tetrahedron (depicted with '$+$' character). The same concerns the extreme values: -1 in edge BC (depicted with 'o' character) and $+1$ in edge AD (depicted with '$*$' character).

3.3. Specialized views of differences between confirmation measures

As the set of available measures is considerable, the practitioners must match measures to particular applications. To guide them in the process, our visualization techniques provide specialized views allowing to identify arguments (i.e. values of a, b, c and d) for which two given measures differ only insignificantly (similarity of the measures) or differ considerably (dissimilarity of the measures). Analogous visualizations may be constructed for groups of measures, with the variance used instead of simple difference.

Consider measures $D(H,E)$ and $M(H,E)$. Figure 6 shows[4] a view of the interior of the difference $D(H,E) - M(H,E)$. Observe that $D(H,E)$ exceeds $M(H,E)$ most in the vicinity of the C vertex, while $M(H,E)$ exceeds $D(H,E)$ most in the vicinity of the B vertex.

4. Conclusions

The paper presents visualization techniques for interestingness measures, which provide practical insights into different details of the analysed measures. The originally 4-dimensional arguments of the measures are effectively represented in 3D using a tetrahedron-based barycentric coordinate system, with values of any operational function, e.g. an interestingness measure, rendered as colour.

The visual analyses are especially useful since they allow to instantly detect and localize interesting characteristics of the measures (extreme values, zeros, etc.), which would otherwise have to be laboriously derived from the analytic definitions of the measures. The implementation of the presented techniques is particularly capable of visualizing: single interestingness measures, regions of interest, i.e. only arguments satisfying pre-defined conditions, differences between pairs of measures or variances of sets of measures. Exemplary applications of the techniques are presented and discussed in detail for a particular group of popular confirmation measures.

5. References

Carnap, R., 1962. *Logical Foundations of Probability, 2nd ed.*. Univ. of Chicago Press.

Christensen, D., 1999. Measuring confirmation. *Journal of Philosophy*, 96:437–461.

Eells, E., 1982. *Rational Decision and Causality*. Cambridge Univ. Press, Cambridge.

Fayyad, U., G.G. Grinstein, and A. Wierse, 2002. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann.

Kemeny, J. and P. Oppenheim, 1952. Degrees of factual support. *Philosophy of Science*, 19:307–324.

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

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

Warren, Joe, 2003. On the uniqueness of barycentric coordinates. In *Contemporary Mathematics, Proceedings of AGGM '02*.

[2] Edges AC and BD for $S(H,E)$ and edges AB, BD and CD for $F(H,E)$ contain undefined values (NaN); the colours used for printing cannot reflect this fact.

[3] The grey colour map is used only to provide the necessary perspective; the colours do not translate to values of the measure (which are constant in this case).

[4] A grey colour map; black and white correspond to the minimum and the maximum of the difference.

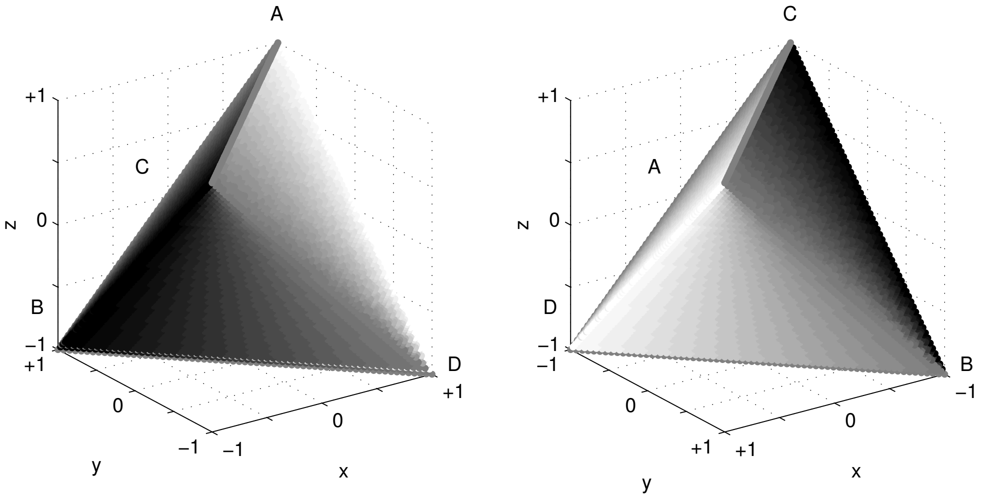

Figure 1: A 3D 2-view regular visualization of $S(H, E)$

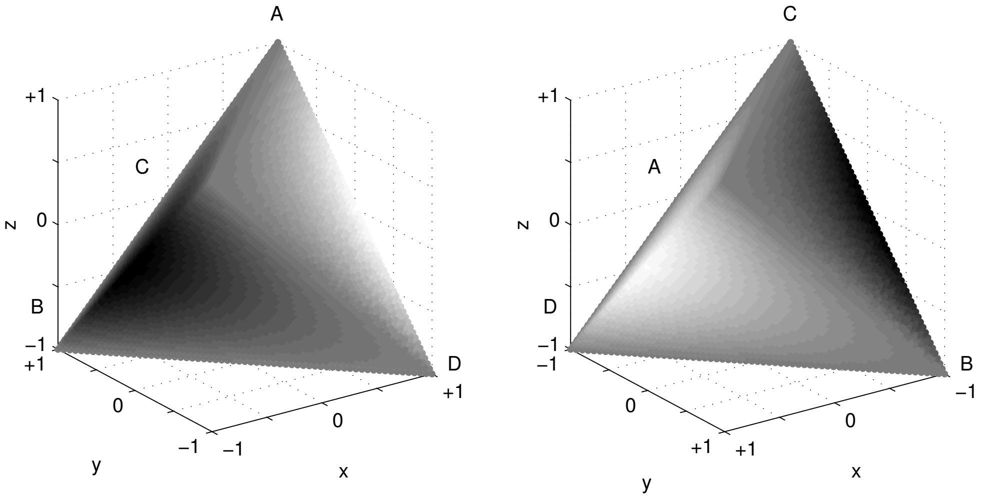

Figure 2: A 3D 2-view regular visualization of $C(H, E)$

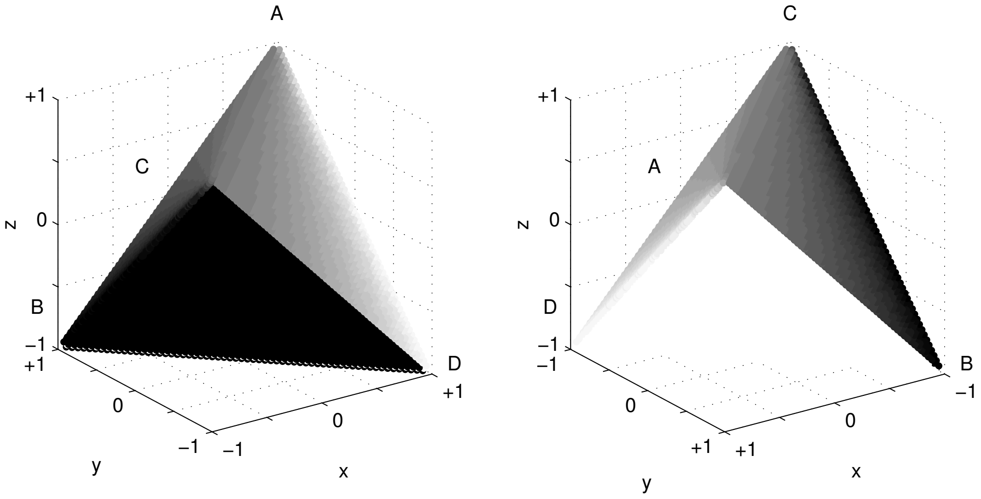

Figure 3: A 3D 2-view regular visualization of $F(H, E)$

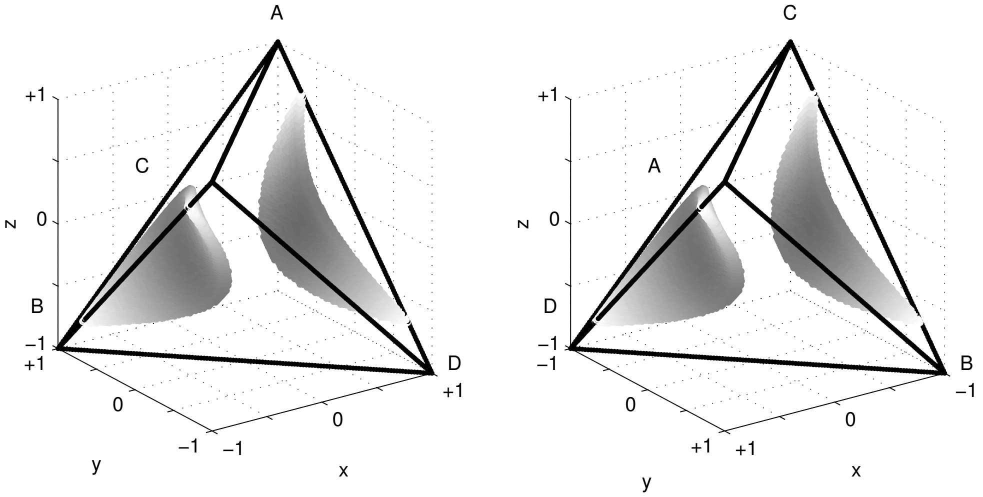

Figure 4: A 3D 2-view specialized visualization of $|C(H,E)| = 0.5$ regions

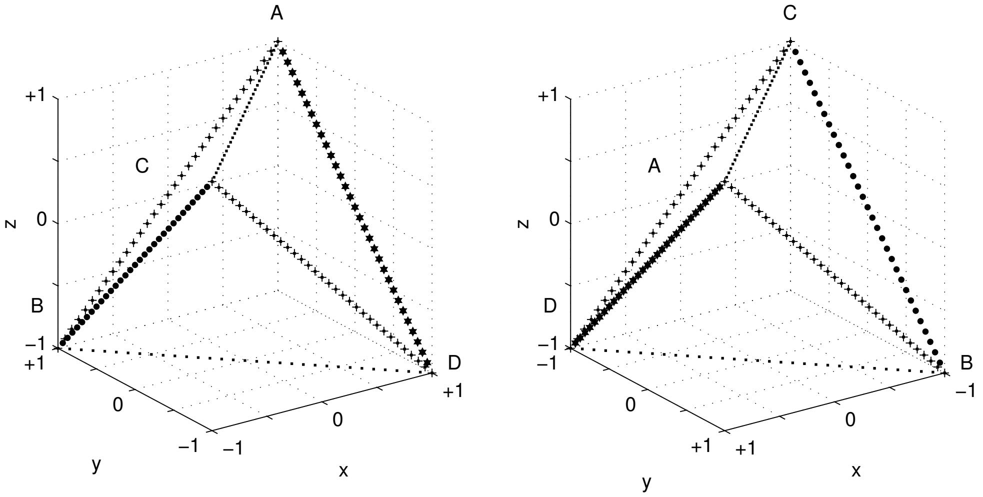

Figure 5: A 3D 2-view specialized visualization of extreme/non-numeric values of $N(H,E)$

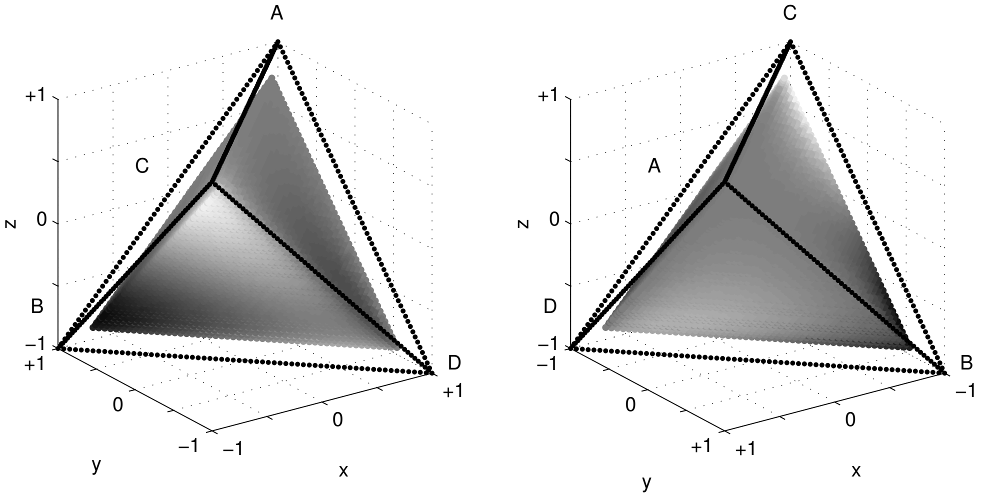

Figure 6: A 3D 2-view specialized visualization of the interior of $D(H,E) - M(H,E)$