An Assertive Will for Seeing and Believing

Introducing a Feature Cardinality Driven Distance Measure to Uninformative Distributions

Joan Garriga

Departament de Sistemes i Llenguatges Informàtics
Universitat Politècnica de Catalunya
jgarriga@lsi.upc.edu

Abstract. What regard should a learning algorithm hold for the different information traces found in a sample? Answering this question objectively is not easy. Moreover, given that a full range of traits can be found in a human learning analogy, from the most daring or ingenuous, to the most conservative or incredulous. But in AI domains it is a must to clearly state the right will for believing what is seen when mining data bases. A key concept in this matter is assertiveness. The aim of this work is to ponder an approach to assertive KDDB, based on a feature cardinality driven distance measure to uninformative distributions. From this perspective, we present an alternative option to the support-confidence framework. The biases of this measure have not yet been thoroughly studied but the measure itself has proved to be quite effective as a heuristic when searching to optimize a sample in a simultaneous multi-interval discretization of continuous features. The empirical results show that the most relevant association or classification rules are revealed. Also, optimal cardinalities and optimal subsets of parents are found for any feature, according to a natural bias toward the MDL principle. As a conclusion, it appears the measure assertively captures knowledge. This may be useful for other data mining issues.

1 Introduction

It is nothing new to point out that some kind of a disappointing shadow of confusion hovers over the data mining scene. The flurry of different measures as well as the comprehensive literature on selecting the right ones for each task at hand ([4],[6]) is no more than a symptom.

In my opinion, three basic objections are the culprit: (i) the stochastic essence of any sample is somewhat misunderstood, (ii) some subtleties about what knowledge is or, more precisely, what better knowledge is, are somewhat set aside, and (iii) the will for believing what is seen is not clearly stated.

These objections are further exposed in the next section as well as throughout this paper. They form the basis for introducing an alternative approach to knowledge discovery wherein a new measure is suggested. The aim is to present

this approach as an open door to further research while the expression given for the measure is yet to be considered an open question.

A thorough analysis on the properties [2] and biases [3] of this measure , as well as some examples should be presented, but unfortunately, space is limited.

In order to state a general framework addressing *association* and *classification rules*, as well as *feature subset selection*, *clustering* and *graphical modelling* issues we will use the following general terminology. Let's consider a domain or concept characterized by a set of m multinomial features $X = \left\{X^1, X^2, \ldots, X^m\right\}$ and a set $\{D\}$ of N examples over these features. Let's consider two any features of this domain and denote $X^p = \{x_1^p, x_2^p, \ldots, x_r^p\}$ and $X^q = \{x_1^q, x_2^q, \ldots, x_s^q\}$ as the set of possible outcomes of features X^p and X^q with cardinalities $crd\left(X^p\right) = r$ and $crd\left(X^q\right) = s$, respectively. Also, for any pair $\left(x_i^p, x_j^q\right)$ we denote n_i^p, n_j^q and n_{ij}^{pq} as the marginal and joint frequencies given in $\{D\}$.

Additionally and for the purpose of clarity, we state three levels of relationship: (i) we refer to a *rule* whenever we are considering a relation like $x_i^p \rightarrow x_j^q$, (ii) we refer to a *subpattern* whenever we are considering the set of rules included in $X^p \rightarrow x_j^q$ or $x_i^p \rightarrow X^q$, and (iii) we refer to a *pattern* whenever we are considering the whole set of rules included in the relation $X^p \rightarrow X^q$. These designations will hold, unless explicitly noted, independently of our intention when considering the relationships (association, classification or whatever).

2 Some Objections to Objective Measures

The most important group of objective measures is based on probability. Given a rule $x_i^p \rightarrow x_j^q$, *coverage* is given as the marginal probabilities of antecedent $P\left(x_i^p\right)$ and consequent $P\left(x_j^q\right)$ of the rule, *support* is given by the joint probability $P\left(x_i^p, x_j^q\right)$, and *confidence* is given by the conditional probability $P\left(x_i^p \left| x_j^q\right.\right)$. Down from here, all objective measures of interestingness combine in different ways these or directly related factors, taken from raw data.

Let's consider the simple example of a transaction data set given in Tab.1. [1]

Milk	Bread	Eggs
1	0	1
1	1	0
1	1	1
1	1	1
0	0	1

Table 1. Transaction Dataset

Coverage for Milk is 4/5 and hence *coverage* for NoMilk is only 1/5. Does it make any sense to consider a rule like Milk → Bread when there is no comparable evidence in the dataset for the rule NoMilk → Bread?

[1] This example is extracted from [2].

Some of the defined measures try to take this fact into account, introducing factors with the probabilities for counter facts in some way. But that is not the question. The real question is whether there is some evidence missing in the dataset in order to adeptly measure the significance of that possible rule. This topic is not new, and has two loose ends:

1. Due to its stochastic nature, any sample should be considered as being less than 100% reliable. Therefore, whenever we consider evidential support from raw data, the estimates we make are afected by the subjective consideration of the sample as being 100% reliable, even though they are estimates. In other words, would it be fair to always estimate a 0/100% of probability for a rule with a 0/100% of *support*?
2. On the other hand, a rule should always be considered, at least, within the framework of its subpattern [5]. If a dependence relationship between two features do exist, this dependence should be patent for the whole pattern. From this point of view, it is important to distinguish between *structural evidence* and *parametrical evidence*. The former relates to the pattern or subpattern levels and expresses whether a remarkable relationship may exist. The latter refers to each one of the rules in the pattern and expresses how this relationship acts whenever it exists.

Let's think again about the transaction example of bread, milk and eggs. For an association rule like Milk → Bread, we have a *support* of 3/5 and a *confidence* of 3/4 and for an association rule like (Milk, Bread) → Eggs we have a *support* of 2/5 and a *confidence* of 2/3. While the combination (Milk, Bread) has a total of four possible outcomes, the combination (Milk, Bread, Eggs) offers as much as eight possible outcomes, therefore with a much lower prior probability. Should we really believe that the former is better supported than the latter? Should we consider these levels of *confidence* from an absolute perspective? In other words, is the same kind, quantity/quality, of knowledge given by these two rules?

In this case, the argument is quite subtle and it has to do with the level of certainty/uncertainty associated with a feature as a function of its cardinality or what is also referred to as the quantity/quality of knowledge given by that feature. The larger the cardinality of the features involved in a rule, the more accurate and valuable is the information, but the lesser the prior probability of finding that rule in the dataset.

These topics have been somewhat overlooked, and this new approach tries to offer a way to address this omission.

3 Assertiveness by means of Objectivity

One really assertive measure should be defined by assuring an impartial comparison within any rule's evidence detected in the sample. Recalling the objections raised above, three conditions should be met for this assumption to be true: (i) the sample should be 100% reliable and equilibrated or otherwise this should be taken into account in some way, (ii) the quantity/quality of knowledge expressed

by the rule should be taken into account in some way, and (iii) the fairest balance between seeing and believing should be guaranteed.

In order to define such and impartial measure we state the following three concepts:

Definition 1. *A feature* $X^p \in X$*, with* $crd\left(X^p\right) = r$*, is in* perfect marginal distribution *(pmd) whenever all its possible outcomes are equally covered, that is,* $\forall\, x_i^p \in X^p$ *all marginal frequencies are* $n_i^p = N/r$

Ideally, if all features in a sample were in *pmd*, all rule's prior probability would be maximally equilibrated.

Definition 2. *Two features* $\left(X^p, X^q\right) \in X$*, with* $crd\left(X^q\right) = s$*, are in* absolutely incoherent conditional distribution *(aicd) whenever* $\forall \left(x_i^p, x_j^q\right) \in \left(X^p, X^q\right)$ *all joint frequencies are* $n_{ij}^{pq} = n_i^p/s$

Again, this is an ideal situation, possible only between features with equal cardinality, but clearly conveys a state of minimum information.

Definition 3. *The* knowledge factor Q*, which is only briefly introduced here, is defined as the degree of accuracy associated to a feature* X^q *as a function of its cardinality,* $crd\left(X^q\right) = s$*, given by,*

$$Q = (s-1)/s \tag{1}$$

On one hand, independently from any sample or domain, *pmd* and *aicd* state two clearly defined uninformative distributions to take distances from:

1. for feature X^q, an expression of its marginal distribution distance to the *pmd* is given by,

$$\Delta\left(X^q\right) = \sum_j \left(\frac{n_j^q - \frac{N}{s}}{\frac{N}{s}} \right)^2 = \sum_j \left(s\,\frac{n_j^q}{N} - 1 \right)^2 \quad . \tag{2}$$

2. respect to feature X^p, an expression of X^q's conditional distribution distance to the *aicd* is given by,

$$\Delta\left(X^q \,|X^p\right) = \sum_{i,j} \left(\frac{n_{ij}^{pq} - \frac{n_i^p}{s}}{\frac{n_i^p}{s}} \right)^2 = \sum_{i,j} \left(s\,\frac{n_{ij}^{pq}}{n_i^p} - 1 \right)^2 \quad . \tag{3}$$

What should be kept in mind, is that expressions (2) and (3) are measuring exactly the same concept.[2]

On the other hand, raw distances given in (2) and (3) are clearly affected by a strong bias due to the cardinality of the features.

[2] Strictly speaking, this expressions don't hold the formal properties of a metric distance functional (particularly, the triangular inequality does not make sense). They should rather be regarded as deviations.I hope this is not going to be misleading.

The philosophy behind this approach is that, taking the *knowledge factor* as a base expressing the quantity/quality of knowledge, a transformation can be applied in order to address this bias. The main contribution of this work is to present a general expression for this transformation, wherein alternative and significantly different measures to *coverage*, *support* and *confidence*, can be derived. These new measures intend to be as objective as possible and intend to state the most assertive will to believe what is seen. The final purpose is to allow an objective comparison between any trace of rule/pattern found, regardless of the actual reliability of the sample and regardless of the cardinality of the features involved, addressing the objections formerly exposed.

4 Defining the Measure

A useful transformation of such distances is given by the general function,

$$Z(x) = exp\left(\alpha \frac{ln(Q)}{Q^2}(s\,x-1)^2\right) \quad , \tag{4}$$

where x can either refer to the marginal or conditional distribution, n_j^q/N or n_{ij}^{pq}/n_i^p, whatever be the case.[3]

Aiming at simplicity, this expression can be rewritten as,

$$Z(x) = bxp\left(\frac{1}{Q^2}(s\,x-1)^2\right) \quad , \tag{5}$$

where *bxp (knowledge factor exponential base)* is a self allowed notation, derived from *exp (natural exponential base)* with analogous meaning, that is, $bxp(K) \equiv Q^K$.

The proximity of this function to a *normal* distribution, $N\left(\frac{1}{s}, \frac{Q}{s}\sqrt{\frac{-1}{2\,ln(Q)}}\right)$ is clear, with two obvious differences which are: (i) it is not a probability distribution, but a distance distribution, so not normalized as a *mass function*, and (ii) it makes sense only in the range $0 \leq x \leq 1$.

Therefore I call this function the *QNormal distance distribution*, $QN\left(\frac{1}{s}, \frac{Q}{s}\right)$, which is depicted in Fig.1 for different values of s.

At the mean, given by $1/s$, its value is 1, and at the boundaries the values are given by,

$$Z_z \equiv Z(0) = bxp\left(\frac{1}{Q^2}\right) \quad ; \quad Z_n \equiv Z(1) = bxp\left(s^2\right) \quad . \tag{6}$$

[3] The factor α has to do with the prior credibility we can give to the sample. Its thorough treatment lies beyond the scope of this work, so let's consider $\alpha = 1$.

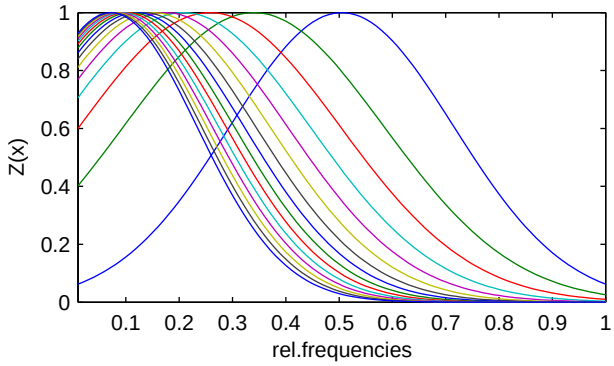

Fig. 1. $QN\left(\frac{1}{s}, \frac{Q}{s}\right)$ for $2 \leq s \leq 15$.

4.1 Presence

Applying the general expression given in (5) to the marginal distribution of feature X^q, we have,

$$\forall x_j^q \in X^q\,, \quad z_j^q \equiv Z\left(\frac{n_j^q}{N}\right) = bxp\left(\frac{1}{Q^2}\left(s\,\frac{n_j^q}{N} - 1\right)^2\right) \quad , \tag{7}$$

Combining (7) with (6) in order to fit values into $(0, 1)$, we can derive an alternative and significantly different measure of *coverage*, which I call *presence*, given by,

$$b_j^q = \frac{1}{s}\left(\frac{z_j^q - Z_z}{1 - Z_z}\right) \;;\; 0 \leq \frac{n_j^q}{N} \leq \frac{1}{s} \quad , \tag{8}$$

$$b_j^q = \frac{1}{s}\left(\frac{z_j^q - Z_n}{1 - Z_n}\right) \;;\; \frac{1}{s} \leq \frac{n_j^q}{N} \leq 1 \quad , \tag{9}$$

This function is depicted in Fig.2. The total *presence* of a feature is then given by $B^q = \sum_j \left(b_j^q\right)$, with a maximum value of 1, given when all possible outcomes for the feature are equally covered. As long as *coverage* of that feature moves away from the *pmd* in any direction, the value of *presence* decreases, vanishing at the boundaries.

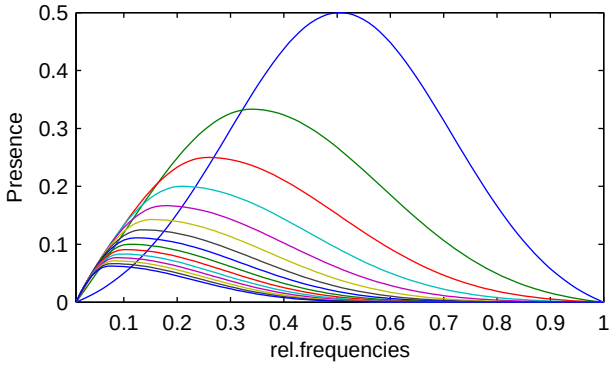

Fig. 2. Presence function for $2 \leq s \leq 15$.

4.2 Coherence

Applying the general expression given in (5) to the conditional distribution $(X^q \mid X^p)$, we have,

$$\forall \left(x_i^p, x_j^q\right) \in (X^p, X^q) \ , \quad z_{ij}^{pq} \equiv Z\left(\frac{n_{ij}^{pq}}{n_i^p}\right) = bxp\left(\frac{1}{Q^2}\left(s\,\frac{n_{ij}^{pq}}{n_i^p} - 1\right)^2\right) \ , \quad (10)$$

Combining (10) with (6) in order to fit values into $(0, 1)$, we can derive an alternative and significantly different measure of *confidence*, which I call *coherence* given by,

$$c_{ij}^{pq} = \frac{1}{r\,s}\left(1 - \frac{z_{ij}^{pq} - Z_z}{1 - Z_z}\right) \ ; \ 0 \leq \frac{n_{ij}^{pq}}{n_i^p} \leq \frac{1}{s} \ , \quad (11)$$

$$c_{ij}^{pq} = \frac{1}{r\,s}\left(1 - \frac{z_{ij}^{pq} - Z_n}{1 - Z_n}\right) \ ; \ \frac{1}{s} \leq \frac{n_{ij}^{pq}}{n_i^p} \leq 1 \ , \quad (12)$$

This function is depicted in Fig.3. The total *coherence* of pattern $X^p \rightarrow X^q$ is then given by $C^{pq} = \sum_{i,j}\left(c_{ij}^{pq}\right)$, with a maximum value of 1, given when each subpattern is maximally coherent, as it is stated in the following definition.

Definition 4. *The conditional distribution* $(X^q \mid X^p)$ *is* maximally coherent *when* $\forall\, x_i^p \in X^p\,, \exists\, x_m^q \in X^q$ *, such that,* $n_{im}^{pq} = n_i^p$ *and* $\forall\, x_{j\neq m}^q \in X^q\,, n_{ij}^{pq} = 0$.

And being both conditions necessary for the maximum *coherence*, they are both assigned the same value of *coherence* $1/\left(r\,s\right)$.

Obviously, it is an asymmetric measure, so that most of the time it will be $c_{ij}^{pq} \neq c_{ji}^{qp}$.

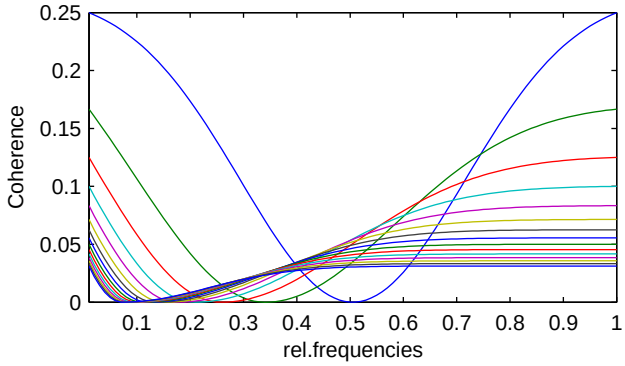

Fig. 3. Coherence function with $r = 2$ and for $2 \leq s \leq 15$.

4.3 Utility

Finally, combining the two former measures, we obtain the *utility* measure for the rule $x_i^p \rightarrow x_j^q$, which is given by,

$$u_{ij}^{pq} = c_{ij}^{pq} \ (b_i^p \, r) \ \left(b_j^q \, s\right) \ , \tag{13}$$

The total *utility* of pattern $X^p \rightarrow X^q$ is then given by $U^{pq} = \sum_{i,j} u_{ij}^{pq}$, with a maximum value of 1, given when *coherence* is maximal and *presence* for both features is perfectly equilibrated.

A depiction example of the *utility* function for $x_i^p \rightarrow x_j^q$ with $(r = 2, s = 3)$ and being X^p in *pmd* is given in Fig.4.

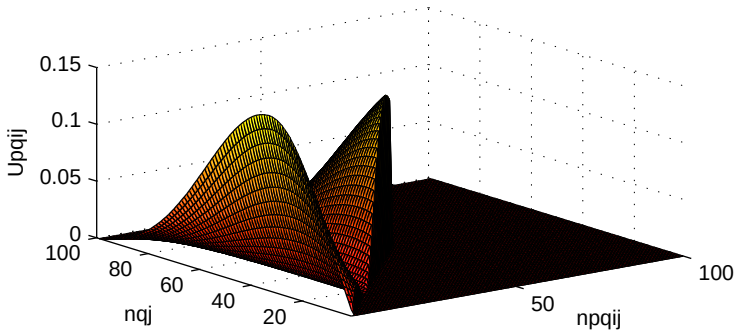

Fig. 4. Utility function for $r = 2$, $s = 3$, $N = 100$, $n_i^p = \frac{N}{r}$

By definition, *utility* is inversely related to the *total amount of uncertainty of the consequent given that the antecedent is known*, (see [1] for a related discussion). Even in the case of *independence*, $U^{p\perp q} \geq 0$, being zero only when X^q is in *pmd*. This expresses the idea that even being independent it is still possible to get some certainty about the consequent, though coming from its own marginal distribution. In such a case, there exists a subspace in the set of all possible joint distributions, in the neighbourhood of *independence*, in which $U^{pq} \leq U^{p\perp q}$. This suggests the daring idea of expanding the concept of independence: it is not the single point where $P\left(X^p, X^q\right) = P\left(X^p\right) P\left(X^q\right)$ but the whole subset of joint distributions for which $U^{pq} \leq U^{p\perp q}$, that is, where the total uncertainty is even greater than that given in *independence*.

4.4 Parametrical Perspective

Finally, the *QNormal distance distribution* holds yet another possible derivation from the parametrical point of view, which clearly explains what it is conceptually being done.

From the inversion of the second half of the curve, we can derive the following expression,

$$\Theta\left(x\right) = \frac{Z\left(x\right)}{2}\,, \quad 0 \leq x \leq \frac{1}{s} \tag{14}$$

$$\Theta\left(x\right) = \left(1 - \frac{Z\left(x\right)}{2}\right)\,, \quad \frac{1}{s} \leq x \leq 1\,. \tag{15}$$

The depiction of this function is given in Fig.5.

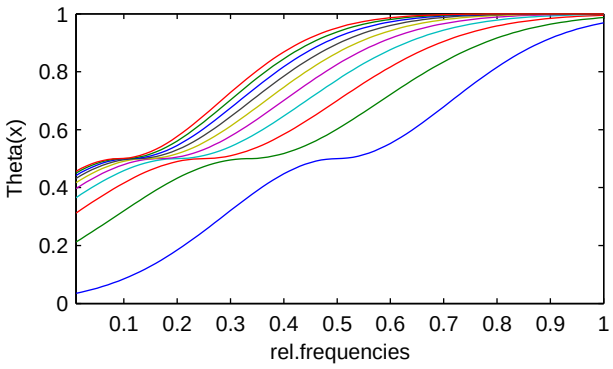

Fig. 5. Theta function for $2 \leq s \leq 15$.

In contrast to the raw interpretation resulting from measures like *coverage*, *support* and *confidence*, this function translates the parameters to a common

space where all of them can be seen relatively to the quantity/quality of knowledge they express.

There's a saddle point at the frequency given by $1/s$, which represents the equilibrium corresponding to the state of minimum information (*pmd* or *aicd*), and moving away from that point this equilibrium is consequently and gradually broken in one or other direction.

The breaking gradient is determined by the σ parameter (depicted in Fig.6), given as,

$$\sigma = \frac{Q}{s} = \frac{(s-1)}{s}\,\frac{1}{s} \quad . \tag{16}$$

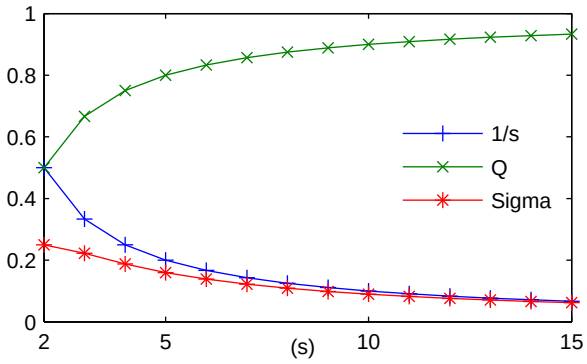

Fig. 6. Sigma function for $2 \leq s \leq 15$.

It combines two factors of s expressing two clashing facts: (i) the Q factor expresses the idea that the more the cardinality, the more accurate the information given by the feature, therefore σ increases and the gradient decreases, so that more evidence must be seen in order to break the equilibrium, (ii) whereas the $1/s$ factor expresses the idea that the more the cardinality, the less the prior probability for the state of both minimum and maximum information (bigger entropy), therefore σ decreases and the gradient increases, making it easier to reach it.

Still another notable difference is that this expression (as depicted in Fig.7) gives non-zero values at the zero frequency and non-one values at the frequency one, therefore providing a straight path to a full family of parameters, that is,

$$\Theta\left(0\right) = \frac{1}{2}\,bxp\left(\frac{1}{Q^2}\right) \quad ; \quad \Theta\left(1\right) = \left(1 - \frac{1}{2}\,bxp\left(s^2\right)\right) . \tag{17}$$

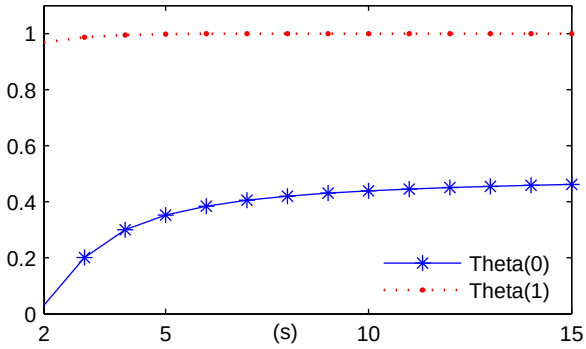

Fig. 7. Theta(0) and Theta(1) functions for $2 \leq s \leq 15$.

The non-zero values express the uncertainty associated to the fact of having no evidence of something. The more the cardinality, the more the prior probability of such a case, so the uncertainty increases. The non-one values express the uncertainty that should be regarded, in spite of having full evidence of something, given the stochastic nature of a sample. The more the cardinality, the less the prior probability of such a case, so the uncertainty decreases and the value tends to one.

Obviously, this expression is not normalized; it is not a *mass function*. It does not directly translate evidence into probabilities; rather, it translates traces of evidence into biases over the equilibrium. Anyway, normalization allows deriving a complete family of parameters from this expression. In classification issues, this conservative understanding of evidence usually turns to be enough and in most of the cases even better than a raw interpretation.

It's hardly worth mentioning, that an interesting option arises from the possibility of applying this parametrical model to any of the measures already existent.

5 Conclusions

This expression intends to give an equable, impartial and equilibrated measure of dependence relationship, taking into account its relative degree of *support* and its associated quantity/quality of knowledge.

Coherence is measured as a trace of dependence. It's to be assumed that whenever two features are dependent, this dependency should be patent for the whole pattern, moving away their conditional distribution from the *aicd*. On the other hand, high rates of *coherence* would be easily achieved with respect to a feature with a great bias in its marginal distribution toward or against one of its possible outcomes. That's the correction introduced into the expression of *utility* by the measure of *presence*. Good *coherence* but poorly or excessively supported

by the sample would be punished by the *presence* factor, giving poor rates of *utility*.

Equanimity is given by the fact that *presence* and *coherence* are measured exactly as the same concept, a distance to their respective uninformative distributions, guarantying this way the most possible assertive balance between seeing (*presence*, *coherence*) and believing (*utility*).

From a summarization point of view, being the measure defined at the least significant level, it can be summed up to whatever may be of interest, providing ranked classifications not only at pattern, subpattern or rule levels, but even at feature and sample levels. Therefore, relevance at each level can be objectively analyzed.

At pattern level, the *utility* measure relates to marginal dependence. However, this measure is directly extensive to relationships like $(X^p, X^q) \rightarrow X^c$. In this case, what is measured turns out to be the relation of conditional dependence $(X^p \perp X^q \mid X^c)$. Therefore, this extended measure of *utility* can be applied to any subset of parents of a feature, providing an ordered list of classification rules. Both matters have significant implications regarding to *clustering* and/or *graphical modelling*. At feature level, conclusions can be derived related to *feature subset selection* issues.

A striking practical application is to implement this measure as a heuristic in a search in order to optimize a simultaneous multi-interval discretization of a sample with some/all continuous features. This application has been tested both in real domain and synthetic data bases, and has shown that the measure leads to optimal cardinalities and optimal subsets of parents for each feature, according to a natural bias to the MDL principle.

References

1. Julien Blanchard, Fabrice Guillet, Henri Briand, and Regis Gras. Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In Proc. of the 11th Int. Symposium on Applied Stochastic Models and Data Analysis (ASMDA'05), 191-200. ENST, 2005.
2. Liqiang Geng, Howard J. Hamilton. Interestingness Measures for Data Mining: A Survey. ACM Computing Surveys, Vol.38, No.3, Article 9, September 2006.
3. Igor Kononenko. On biases in estimating multi-valued attributes. In Proc. of the Fourteenth Int. Joint Conference on Artificial Intelligence (IJCAI'95), 1034-1040, Montreal, Canada, 1995.
4. Philippe Lenca, Patrick Meyer, Benoît Vaillant, Stephan Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. European Journal of Operational Research. Vol.184., No.2, 610-626, January 2008.
5. Alex Tze Hiang Sim, Maria Indrawan, Bala Srinivasan. The importance of negative associations and the discovery of association rule pairs. International Journal of Business Intelligence and Data Mining, Vol.3 No.2, 158-176, September 2008.
6. Pang-Ning Tan, Vipin Kumar, Jaideep Srivastava. Selecting the right interestingness measure for association patterns. In Proc. of the 8th Int. Conference on Knowledge Discovery and data Mining (KDD'02), 32-41, Edmonton, Canada, 2002.