PhD Thesis Report 1209

Joan Garriga
Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

November 26, 2009

Abstract

This report is a quick review about the most up to date state of my research.

Finally, I have the hope of having reached a right general measure of the knowledge conveyed by a distribution. This measure is supported by an axiomatic approach and has a clear interpretation in terms of a geometric representation of knowledge. It applies equally to marginal and conditional distributions, and provides the relation between them and the joint distribution. It also allows for a particular definition of dependence between attributes.

Applying the same axiomatization to disjoint dependent events (in our context, each of the outcomes of any feature), gives rise to the question of the algebraic structure of knowledge, which still remains somewhat confused.

From this general expression of knowledge it is straightforward to give precise definitions of the concepts of *presence* (representativity), *coherence* (reliability) and *utility* (quality), that I have been handling since the very beginning of my work.

Amazingly, after turning the matter over and over, I have ended up with an expression of knowledge, (or certainty), which is very similar to *normalized entropy*, (or uncertainty). This can be easily explained considering that certainty and uncertainty are in essence the same thing, just different degrees of believe.

It turns out that, while being entropy a logarithmic additive measure, knowledge appears to be a multiplicative quadratic measure. Taking into account this difference, a lot of analogies can be observed between both measures. In particular, the set of properties of entropy turns, consequently, to the same set of properties for knowledge: monotonicity, normalization, symmetry, expandability, and analog properties to additivity (for independent attributes) and subadditivity, but referred both to the operation of multiplication.

Our first empirical results show that this measure of knowledge, definitely outperforms entropy. An extended version of this report will be soon ready, including a detailed empirical comparison between both measures. I expect to include also a new section, explaining how the parametrical evidence function fits in this new approach.

Contents

Chapter 1

Structural Evidence

Our intention is to define a measure of the *quality* of the *explanation model* represented by the implication $X^p \to X^q$ (where X^p, X^q can be features or sets of features).

Our initial hypothesis is that the quality of the information conveyed by this implication is a combination of the knowledge explicit by its conditional distribution, and the knowledge implicit in the marginal distribution of both, antecedent and consequent. This two sources of knowledge are clashing in the sense that, the more biased the marginal distribution is, the more predictable is the possible outcome, independently of the fact of considering the possible relation of dependence between them.

Consequently, in order to adeptly measure the quality of an implication we should compose these two kinds of knowledge in a single measure. This measure should express a delicate equilibrium in the way we are combining them, and this is the way the bias/variance dilemma takes form in our context.

A second hypothesis is that the *quality* (richness) of any *explanation model* is going to be definitely related with the cardinality of both, antecedent and consequent of the implication. The way this richness is expressed, is through the uncertainty $U = 1/s$ and certainty $C = (s-1)/s$ factors associated to any cardinality s.

1.1 Axiomatic approach

What we are concerned about is useful, or actionable, knowledge, that is, *knowledge of a consequent given that an antecedent is known.*

In [3], after reviewing Kolmogorov's axiomatization in the context of finite sample spaces, we conclude that *knowledge* $\neq$ *probabilities*. Knowledge is expressed in the form of rules and probabilities, but they are not exactly the same thing. Consequently, probability axiomatization may not be the most appropriate in our context.

The main differences with respect to probabilities refer in particular to the following:

- knowledge can never be zero (in the worst case, we should have equal expectations for each possible outcome, and this states a minimum knowledge greater then zero), neither can be one (absolute knowledge is unfeasible due to the inherent uncertainty of the context);
- knowledge is inherently related to the cardinality of features;
- knowledge should be contrastable.

Hereof, we give the following primary axiomatic approximation to knowledge:

Definition 1. Given a sample space $\Omega = \{X^p, X^q\}$, and given a sample $\mathcal{D}$ drawn from Ω, we write $\mathcal{S} \subset \Omega$ as the set of all events observed in the sample, and its complementary, $\mathcal{U} \subset \Omega$, as the set of all events not observed in the sample. Then, $K(A)$ is a *knowledge* measure over $\Omega = \{\mathcal{S} \cup \mathcal{U}\}$ if:

- Axiom 1. for any event $A \in \Omega$, $U_A \leq K(A) < 1$, being U_A the uncertainty factor associated to A, given at the point of equiprobability.

with analogous definitions for conditional knowledge and independence,

- Axiom 2. for any two events $(A, B) \in \Omega$, $U_B \leq K(B\,|A) < 1$
- Axiom 3. B is independent of A as long as $K(B\,|A) \leq K(B)$

Note that some significant differences with respect to probability axioms are implicit in this axiomatization:

- we are not considering herein any particular algebraic structure for knowledge about disjoint events, therefore no notion of countable additivity is defined;
- a relation of dependence is not an on/off switch: it would be reasonable to consider independence in the range stated by axiom 3, and consider higher degrees of dependence as far as $K(B\,|A) \geq K(B)$.

Also, note that we write $\Omega = \{\mathcal{S} \cup \mathcal{U}\}$ in order to explicitly include the uncertainty inherently associated to any data mining process. What we pretend, is to relax the assumption of *identical distribution*. But, it is not that we are expecting that new examples may come from a possibly diferent distribution, (in this case, our effort to model it would be sensless). Rather, we are cautious that the sample may not convey enough evidence about the whole of the distribution, specially when we are considering a joint explanation model involving many features.

1.2 States of minimum information

In order to measure the knowledge conveyed by a distribution, we define a reference uninformative distribution from which we can take measures of deviation. Our interpretation is that, the larger the deviation, the greater the amount of knowledge expressed by that distribution.

This uninformative distribution is analogously defined for marginal and conditional distributions.

1.2.1 The null conditional distribution (*ncd*)

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

1.2.2 The perfect marginal distribution (*pmd*)

Definition 3. A feature $X^p \in X$, with $|X^p| = r$, 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 $n_i^p = N/r$

It is important to note that, these reference uninformative distributions are independent of any sample or domain under consideration.

1.3 Deviation from minimum information

Given two features X^p with $|X^p| = r$, and X^q with $|X^q| = s$, a reasonable expression of the deviation of their conditional distribution $(X^q \,|X^p)$ with respect to the *ncd*, is,

$$\Delta \left(X^q \,|X^p\right) = \sum_{i,j}^{r,s} \left(\delta_{ij}^{pq}\right)^2 = \sum_{i,j}^{r,s} \left(\frac{n_{i,j}^{p,q}}{n_i^p} - \frac{1}{s}\right)^2 \tag{1.1}$$

A simple illustration of our analogy between marginal and conditional knowledge is to consider the pattern $\oslash \rightarrow X^q$, or better said, $D \rightarrow X^q$, being D the sample data. This relation conveys the prior knowledge about the feature, as it should be inferred from the data, and would be given by the deviation of its marginal distribution with respect to the *pmd*, that is,

$$\Delta \left(X^q\right) = \sum_{j}^{s} \left(\frac{n_j^q}{N} - \frac{1}{s}\right)^2 \tag{1.2}$$

1.4 Geometric interpretation

Let's take as a reference the marginal distribution of feature X^q and its deviation from the *pmd*. The raw deviation of any x_j^q is given by,

$$\delta_j^q = \delta\left(x_j^q\right) = \left(\frac{n_j^q}{N} - \frac{1}{s}\right)$$

Let's fix a square with an area equal to one and let's imagine that this area represents the absolute knowledge. Let's divide each side at the point corresponding to $1/s$, so that we get two portions, according to the *certainty* and *uncertainty* factors. We will refer to the crossing point as the point of minimum information.

Now, let's represent the square of a positive δ_+ and a negative δ_- deviations with respect to the point of minimum information, as it is shown in fig. 1.1

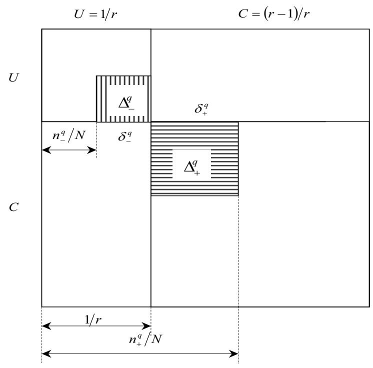

Figure 1.1: Graphical representation of knowledge

It can be observed that the square of the deviations are areas relative to the full square, and we have a graphical representation of knowledge as areas.

A further illustration of how deviations and square deviations are related in our graphical representation of knowledge is given in fig.1.2.

We can easily observe the following basic properties:

- for some values of j we will have positive deviations and for others we will have negative deviations, and they all sum up to zero,

$$\sum_j^s \delta_j^q = \sum_j^s \left(\frac{n_j^q}{N} - \frac{1}{s}\right) = \frac{1}{N}\sum_j^s n_j^q - 1 = 0 \tag{1.3}$$

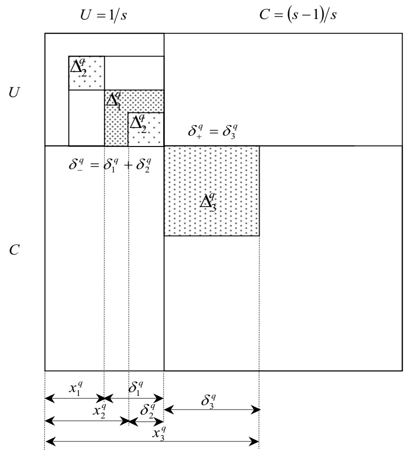

Figure 1.2: Knowledge representation for X^q with $s = 3$

- obviously, this does not hold for square deviations, where the square of the sum is not the sum of the squares, but we have,

$$\sum_{j}^{s}\left(\delta_j^q\right)^2 = \sum_{j}^{s}\left(\frac{n_j^q}{N} - \frac{1}{s}\right)^2 = \sum_{j}^{s}\left(\frac{n_j^q}{N}\right)^2 - U_s \tag{1.4}$$

Hence, when we have maximum deviation, *i.e.* $\sum_j^s \left(\delta_j^q\right)^2 = 1 - U_s$, we still have a lack of knowledge amounting U_s, and what we get is the shadowed areas shown in figure 1.3 for different values of s.

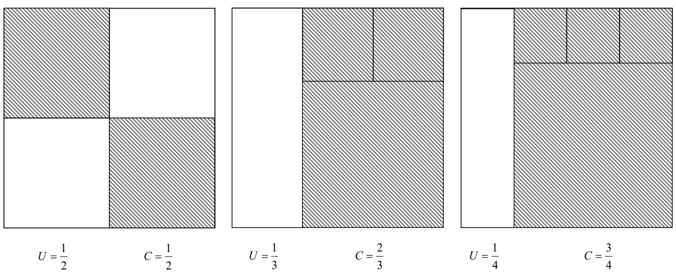

Figure 1.3: Areas of certainty and uncertainty for $s = 2, 3, 4$

For each cardinality, the shadowed, and not shadowed, areas represent the relation between certainty and uncertainty with respect to the absolute knowledge given by the full square. Absolute knowledge, or absence of uncertainty, would only be achieved with an infinite cardinality, that is, a continuous feature.

At the same time, let's note how the relation between quality of knowledge and cardinality, appears naturally from the simple fact of taking measures of deviation with respect to minimum information.

But, does this scaling of knowledge make any sense? Does even make any sense to consider that an additive combination of such areas is certainly related to the global knowledge it should express?

1.5 Marginal knowledge

Referring to our axiomatic approach, the minimum knowledge we can have is that given by the uncertainty factor. That is, at the point of minimum information we have U, and it increases as the square deviations increase.

The most direct expression of this idea is,

$$K\left(X^{q}\right)=U_{s}+\sum_{j}^{s}\left(\delta_{j}^{q}\right)^{2} \tag{1.5}$$

This is intuitive not only from the axiomatic point of view, but also from the geometric point of view, where it is clear that U_s is just the complementary portion of knowledge to get the full square.

In fact, by the property 1.4, the above expression 1.5, is nothing more than,

$$K\left(X^{q}\right)=\sum_{j}^{s}\left(\frac{n_{j}^{q}}{N}\right)^{2}$$

and it is straightforward to show that this measure holds the following properties:

1. normalization
2. monotonicity (with respect to deviation)
3. symmetry
4. expansibility

1.6 Conditional knowledge

The same reasoning applied to marginal distributions in 1.5, can be applied to conditional distributions, where we get,

$$K\left(X^{q}\left|x_{i}^{p}\right.\right)=U_{s}+\sum_{j}^{s}\left(\delta_{ij}^{pq}\right)^{2} \tag{1.6}$$

which equally simplifies to,

$$K\left(X^q \left| x_i^p \right.\right) = \sum_j^s \left(\frac{n_{ij}^{pq}}{n_i^p} \right)^2$$

As desirable, in case of independence, this expression yields,

$$K\left(X^q \left| x_i^p \right.\right) = K\left(X^q\right)$$

This gives a clear interpretation of how the knowledge conveyed by the conditional distribution relates to the dependence relation between antecedent and consequent:

- given a particular evidence, the minimum knowledge we can have is $K\left(X^q \left| x_e^p \right.\right) = U_s$, which corresponds to the point of equiprobability or *equilibrium*;
- at the point of independence we will have $K\left(X^q \left| x_e^p \right.\right) = K\left(X^q\right) \geq U_s$;
- from that point on, $K\left(X^q \left| x_e^p \right.\right) > K\left(X^q\right)$ and we may begin to consider a possible relation of dependence;
- and in case of absolute dependence, $K\left(X^q \left| x_e^p \right.\right) = 1$

For the whole pattern $X^p \rightarrow X^q$, it is going to be useful to define $K\left(X^q \left| X^p \right.\right)$ in terms of a weighted mean expression of the knowledge conveyed by the conditional distribution given each x_i^p, that is,

$$K\left(X^q \left| X^p \right.\right) = \sum_i^r \frac{n_i^p}{N} K\left(X^q \left| x_i^p \right.\right) = U_s + \sum_i^r \frac{n_i^p}{N} \sum_j^s \left(\delta_{ij}^{pq}\right)^2 \tag{1.7}$$

1.7 Joint distributions

Following from the general expression, we have,

$$K\left(X^p, X^q\right) = U_{rs} + \sum_{i,j}^{r,s} \left(\frac{n_{ij}^{pq}}{N} - U_{rs} \right)^2 = \sum_{i,j}^{r,s} \left(\frac{n_{ij}^{pq}}{N} \right)^2 \tag{1.8}$$

In this case, we have the following relation between the knowledge conveyed by the joint distribution and the dependence relation between antecedent and consequent:

- The minimum knowledge we can have is $K\left(X^p, X^q\right) = U_{rs}$, which corresponds to the point of equiprobability or *equilibrium*.

- In case of independence we have,

$$\sum_{i,j}^{r,s}\left(\frac{n_{ij}^{pq}}{N}\right)^2=\sum_{i}^{r}\left(\frac{n_{i}^{p}}{N}\right)^2\sum_{j}^{s}\left(\frac{n_{j}^{q}}{N}\right)^2$$

thus,

$$K\left(X^p,X^q\right)=K\left(X^p\right)\,K\left(X^q\right)\geq U_{rs} \tag{1.9}$$

Let's note that this is an analog property to *additivity*, but with respect to *multiplication*.

- From that point on, we may begin to consider a relation of dependence, and following from,

$$\sum_{i,j}^{r,s}\left(\frac{n_{ij}^{pq}}{N}\right)^2=\sum_{i}^{r}\left(\frac{n_{i}^{p}}{N}\right)^2\sum_{j}^{s}\left(\frac{n_{ij}^{pq}}{n_{i}^{p}}\right)^2$$

we get,

$$K\left(X^p,X^q\right)=\sum_{i}^{r}\left(\frac{n_{i}^{p}}{N}\right)^2\,K\left(X^q\left|x_i^p\right.\right) \tag{1.10}$$

So, what we have is a particular composition of the marginal and conditional knowledge given each particular outcome of the antecedent.

Because of dependence, $\forall x_i^p$, $K\left(X^q\left|x_i^p\right.\right)\geq K\left(X^q\right)$, therefore,

$$\begin{aligned}K\left(X^p,X^q\right)&=\sum_{i}^{r}\left(\frac{n_{i}^{p}}{N}\right)^2\,K\left(X^q\left|x_i^p\right.\right)\\&\geq\sum_{i}^{r}\left(\frac{n_{i}^{p}}{N}\right)^2\,K\left(X^q\right)=K\left(X^p\right)\,K\left(X^q\right)\end{aligned} \tag{1.11}$$

So, the measure holds also an analog to the *subadditivity* property, but in the opposite direction, and expressed again, in terms of multiplication.

- And in case of absolute dependence, for each x_i^p, $K\left(X^q\left|x_i^p\right.\right)=1$, and consequently,

$$K\left(X^p,X^q\right)=K\left(X^p\right) \tag{1.12}$$

1.8 Disjoint dependent events

At this point we are coming across with the most controversial question of our approach. Being consequent with our axiomatization, the knowledge we can have with respect to each elementary event x_j^q, should be given analogously by,

$$K\left(x_j^q\right) = U_s + \left(\delta_j^q\right)^2 = U_s + \left(\frac{n_j^q}{N} - U_s\right)^2 \tag{1.13}$$

In some way, this should be related with the portion of knowledge that each element contributes to $K\left(X^q\right)$. But it is obvious that, nor $K\left(X^q\right) = \sum_j^s K\left(x_j^q\right)$, neither $K\left(X^q\right) = \prod_j^s K\left(x_j^q\right)$. Thus, the question about the algebraic structure of knowledge remains somewhat confused. How should we think about knowledge when it looks not to be additive neither multiplicative?

One alternative is to redefine eq. 1.13 as,

$$K\left(x_j^q\right) = U_s^2 + \left(\delta_j^q\right)^2 = U_s^2 + \left(\frac{n_j^q}{N} - U_s\right)^2$$

in which case we would certainly have $K\left(X^q\right) = \sum_j^s K\left(x_j^q\right)$.

But it would not be quite fair to justify that uncertainty with respect to any outcome x_j^q is reduced to U_s^2 in case of minimum information.

Moreover, the relations given in 1.9 and 1.10 suggest that knowledge should combine by multiplication rather than addition. This is the reason why I posed that bridge to information theory [3], in which I proposed an exponential based structure of knowledge. But this posing comes along with some new controversial questions and I'm still working on it.

1.9 Functions of structural evidence

Once we have given a general definition of the knowledge conveyed by marginal and conditional distributions, it is straightforward to derive the measures of structural evidence of any implication $X^p \to X^q$.

1.9.1 Coherence

Coherence is a measure of the *reliability* of the whole implication involved by the pattern $X^p \to X^q$, and is given by the mean conditional knowledge given by eq 1.7,

$$C^{pq} = K\left(X^q \left| X^p\right.\right) = U_s + \sum_i^r \frac{n_i^p}{N} \sum_j^s \left(\delta_{ij}^{pq}\right)^2 \tag{1.14}$$

1.9.2 Presence

Presence is a measure of the *representativity* of the sample with respect to a feature, and is given by the opposite of the marginal knowledge, that is,

$$B^q = 1 - K\left(X^q\right) = C_s - \sum_j^s \left(\delta_j^q\right)^2 \tag{1.15}$$

1.9.3 Utility

Utility is a composition of the two formers, in which both clashing sources of knowledge are combined, in order to get a measure of the quality of the implication $X^p \rightarrow X^q$, as an *explanation model* of X^q,

$$U^{pq} = C^{pq}\, B^q \tag{1.16}$$

Note that whenever X^q is in *pmd*, we have $B^q = C_s$. That means that, in the best case, our knowledge about X^q would be $U^{pq} = C_s\, C^{pq}$.

In other words, *utility* is indeed the right upper bounded expression of knowledge, in compliance with our first axiomatization [3], that we had already introduced in eq. 3.9 in [3], that is,

$$K\left(X^q\right) = C_s\left(U_s + \sum_j^s \left(\delta_{ij}^{pq}\right)^2\right)$$

1.10 Knowledge and Entropy

At that point, it can be hardly overlooked, that after turning the matter over and over, we have ended up with a very simple expression (far from our first posing), which is closely related to *entropy*[1], holding analog definitions and analog properties. This is shown in table 1.1.

Therefore, because of the close (though reversed) conceptual meaning of both measures, *presence* and *coherence* could be directly defined in terms of *normalized entropy*, that is,

$$B^q = H_n\left(X^q\right)$$

$$C^{pq} = 1 - H_n\left(X^q \left| X^p\right.\right)$$

and composed in an expression of *utility*, exactly in the same way that in eq. 1.16.

[1]In order to get similar values we should consider *normalized entropy*

Entropy	Knowledge
$H\left(X^q\right) = -\sum_j^s \frac{n_j^q}{N} log \frac{n_j^q}{N}$ $0 \leq H\left(X^q\right) \leq log\left(s\right)$	$K\left(X^q\right) = \sum_j^s \left(\frac{n_j^q}{N}\right)^2$ $U_s \leq K\left(X^q\right) < 1$
$H\left(x_j^q\right) = -\frac{n_j^q}{N} log \frac{n_j^q}{N} \quad (!?)$	$K\left(x_j^q\right) = U_s + \left(\frac{n_j^q}{N} - U_s\right)^2$
$H\left(X^q \mid x_i^p\right) = -\sum_j^s \frac{n_{ij}^{pq}}{n_i^p} log \frac{n_{ij}^{pq}}{n_i^p}$ $H\left(X^q \mid X^p\right) = -\sum_i^r \frac{n_i^p}{N} H\left(X^q \mid x_i^p\right)$ $0 \leq H\left(X^q \mid X^p\right) \leq H\left(X^q\right)$	$K\left(X^q \mid x_i^p\right) = \sum_j^s \left(\frac{n_{ij}^{pq}}{n_i^p}\right)^2$ $K\left(X^q \mid X^p\right) = \sum_i^r \frac{n_i^p}{N} K\left(X^q \mid x_i^p\right)$ for $X^q \perp X^p$ $U_s \leq K\left(X^q \mid X^p\right) \leq K\left(X^q\right)$ otherwise $K\left(X^q\right) < K\left(X^q \mid X^p\right) < 1$
$H\left(X^p, X^q\right) = -\sum_{i,j}^{r,s} \frac{n_{ij}^{pq}}{N} log \frac{n_{ij}^{pq}}{N}$ for $X^q \perp X^p$ $H\left(X^p, X^q\right) = H\left(X^p\right) + H\left(X^q\right)$ otherwise $0 < H\left(X^p, X^q\right) < H\left(X^p\right) + H\left(X^q\right)$	$K\left(X^p, X^q\right) = \sum_{i,j}^{r,s} \left(\frac{n_{ij}^{pq}}{N}\right)^2$ for $X^q \perp X^p$ $U_{rs} \leq K\left(X^p, X^q\right) \leq K\left(X^p\right) K\left(X^q\right)$ otherwise $K\left(X^p\right) K\left(X^q\right) < K\left(X^p, X^q\right) \leq K\left(X^p\right)$

Table 1.1: Definitions and properties.

So the question is now, was it worth, after all? Did not already exist such a measure? And in case it does not, what new contributions may we get from this approach? Answering to these questions is the job in which I'm working now.

Let's forward that the key point is that while entropy is defined in terms of *uncertainty* and *minimum description length*, our measure is directly defined in terms of *knowledge*, (certainty), and *maximum quality* expression of it, which in turn, leads equally to minimum description length models. The advantage is that, from our side, we are giving a more direct and clear interpretation of knowledge.

I have performed some empirical comparisons, using the synthetic toy domain already presented in [2] and [3]. Beyond all doubt, the conclusion that can be drawn from the results, is that the measure of *knowledge* performs much better then *entropy*: convergence to the right model is more reliable and faster, and robustness against bias and sample size is definitely better.

This better performance can be explained basically from three facts: (i) *knowledge* is more unbiased with respect to cardinality; (ii) the delicate balance between *coherence* and *presence*, (our expression of the bias\variance dilemma), is better achieved with our measure, (in the way I like it, it states *the right will for seeing and believing*); and (iii) *knowledge* has a preference for lower cardinalities whenever evidence is blurred, and has a preference for higher cardinalities whenever the representativity of the sample is good enough, while *entropy* is neutral with respect to these questions. This can be seen in fig. 1.4.

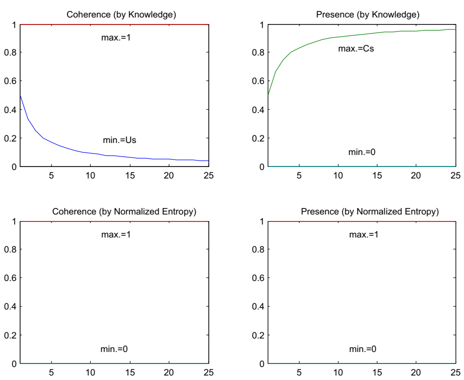

Figure 1.4: Maximum and minimum values for $2 \leq s \leq 25$

This special combination of bias toward simple models in the measure of *coherence*, and bias toward complex models in the measure of *presence*, is what leads to the highest quality models.

1.11 Concluding

We have finally achieved an expression that has a nice geometric interpretation in terms of knowledge conveyed by a distribution. At the same time, this expression holds the basic properties that one would expect from an axiomatic point of view.

But, one of the most important, is how this measure expresses that, the notion of *quality of knowledge* appears naturally related to the cardinality of the features, and it is inherent to the simple fact of measuring deviations with respect to the minimum information.

This does not state yet any guarantee that knowledge should follow this behavior. But indeed, it is an interesting indication that our geometric representation of knowledge may express by itself a right cardinality scaling of knowledge.

The empirical validations suggest that, at least for this particular kind of problem, our measure definitely outperforms entropy. In this case, we use our measure as a heuristic in order to search within the whole space of models that may describe the domain, and in fact, it leads us to the right one. Hence, we are ranking the right rules as the best. My reasoning is that, conversely, when facing a different kind of problem in which the set of rules is already defined, the rank given by our measure should place the best ones in the first positions. Therefore, it is reasonable to expect that this measure may be useful for other data mining issues.

Finally, let's point out the following suggestive consideration: being our measure a descriptive measure, it has its roots close to a statistical measure as it is the χ^2, and has a formal definition and properties close to entropy based considerations. That is, in some sense, it embodies some of the characteristics of all of them, being placed somewhere in the middle of the space of interestingness measures.

Chapter 2

Parametrical Evidence

When talking about *parametrical evidence*, we refer to the fact of observing the frequencies present in the sample, with the aim of learning plausible parametric models that can be associated to the implication $X^p \rightarrow X^q$, namely, the marginal distribution of the antecedent, $P(X^p)$, and the conditional distribution of the consequent, $P(X^q | X^p)$.

Every learning paradigm in knowledge discovery,

What is our the difference between structural and parametrical evidence?

The second hypothesis expresses the idea that the global evidence conveyed by a sample, refers to two different aspects about the represented domain: (i) the first one relates to *which* are the relations among features that actively operate in the behaviour of the domain, what is known as the *dependencies model*, and (ii), the second one relates to *how* this active relations operate in case they certainly exist. We refer to the former as *structural evidence*, which will determine the topology of the model, and to the latter as *parametrical evidence*, which will determine the parameters of the model.

It is clear to our intuition, that whenever a dependence exist among two features, it should be independent of any unbalance in the marginal distributions. This will affect their parametrical relation but not the dependence nature of its relation.

From this consideration, knowledge discovery can be tackled as two separate problems, one related to the structure of the model and another one related to the parameters of the model, (in association rule mining, the optimal set of rules and the confidence of the rules).

2.1 Axiomatization

Bibliography

[1] Garriga J. An Assertive Will for Seeing and Believing. Introducing a Feature Cardinality Driven Distance Measure to Uninformative Distributions. In *Proceedings of the Quality issues, measures of interestingness and evaluation of data mining models (QIMIE'09)* International Workshop, Bangkok, April 2009.

[2] Garriga J. The Right Will for Seeing and Believing. Recently presented PhD Thesis project. June 2009.

[3] Garriga J. Groundwork for a new approach to knowledge discovery. September 2009.