Groundwork for a New Approach to Knowledge Discovery

Joan Garriga
Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

October 15, 2009

Abstract

We present some theoretical background and recent improvements about our feature cardinality driven distance measure to uninformative distributions.

Contents

1 Introduction **2**

2 Axiomatization **3**

2.1 Kolmogorov's Axioms 3

2.2 Probability Axioms 4

2.3 Knowledge Discovery Axioms 5

2.4 The quality of knowledge 7

3 Geometry **9**

3.1 Deviation from minimum information 9

3.2 Graphical representation of knowledge 10

3.3 Half dependence 17

4 Information theory **18**

4.1 Presence and Coherence 19

4.2 The scaling and shaping factors 20

4.3 Half dependence revisited 22

5 Conclusions **23**

5.1 Examples 24

Bibliography **27**

A **28**

A.1 Cardinality scaling of knowledge 28

A.1.1 Non-subsequent cardinalities 29

A.1.2 Arbitrary distributions 30

Chapter 1

Introduction

This work is an uncomplete, six weeks deadline, effort to present some theoretical background to our recently introduced new approach to knowledge discovery, (please refer to [2], [3]).

Some new ideas are presented:

- In the second chapter we present an axiomatic approach to knowledge discovery. Kolmogorov's axiomatization, and its simplified probability version, are reviewed in order to state the relation of knowledge with probability. Although being knowledge closely related with frequencies, our conclusion is that probability axiomatization may not be the most appropriate in the context of knowledge discovery. Therefore we present an alternative 'list of axioms', (let's say it), for this context.

- In the third chapter we present a geometric interpretation of our measure of deviation from uninformative distributions. This perspective suggests a graphical representation of knowledge from which many interesting new ideas can be derived. The most important one is that, contrary to what we had previously presented, it turns out that the measure may state by itself a right cardinality scaling of knowledge.

- In the fourth chapter we present yet a different approach, setting a bridge to information theory and contributing further theoretical support to our work.

- Finally, we end with a short statement of our up to date conclusions and future work. We also present some examples showing that, with respect to our previous approach, better results can be achieved with this one.

Chapter 2

Axiomatization

2.1 Kolmogorov's Axioms

The axiomatic approach to probability, as originally formulated by Kolmogorov in 1933, states verbatim [5]:

Let $\mathcal{E}$ be a collection of elements ... which we shall call *elementary events*, and $\mathcal{F}$ a set of subsets of $\mathcal{E}$; the elements of the set $\mathcal{F}$ will be called *random events*.

- Axiom 1. $\mathcal{F}$ is a field of sets.
- Axiom 2. $\mathcal{F}$ contains the set $\mathcal{E}$.
- Axiom 3. To each set A in $\mathcal{F}$ is assigned a non-negative real number $P(A)$. This number $P(A)$ is called the probability of the event A.
- Axiom 4. $P(\mathcal{E}) = 1$.
- Axiom 5. If A and B have no element in common, then

$$P(A+B) = P(A) + P(B)\ . \tag{2.1}$$

About axiom 5, let's note that Kolmogorov simply uses $+$, instead of the $\cup$ symbol, and explicitly states the condition that A and B are disjoint instead of the expression $A \cap B = \oslash$.

Kolmogorov adds that, a system of sets is a field, if it also contains the sum, difference and product of any two of its sets. In modern notation, for any $A \in \mathcal{F} \to A^c \in \mathcal{F}$, and we can derive that $A \cap A^c = \oslash \in \mathcal{F}$ and $A \cup A^c = \mathcal{E} \in \mathcal{F}$, but it is not necessary to postulate $P(\oslash) = 0$ or $P(A) \leq 1$, because it is all implicit in the above axioms.

If $\mathcal{E}$ is an infinite collection of elements, then $\mathcal{F}$ is normally restricted to be such that it is closed under countable unions of sets, and axiom 5 is replaced by the σ-additivity condition,

- Axiom 5'. If A_n is a set of pairwise disjoint sets in $\mathcal{F}$, then

$$P\left(\cup_n A_n\right) = \sum_n P\left(A_n\right) \ . \tag{2.2}$$

This axiomatization is complemented with the notions of *stochastic independence* and *conditional probability*:

- Axiom 6. The *necessary* and *sufficient* condition that A and B be stochastically independent events is,

  $$P\left(A \cap B\right) = P\left(A\right) P\left(B\right) \ . \tag{2.3}$$

  Note that this is not always equivalent to *physical* independence.

- Axiom 7. If $P\left(B\right) \neq 0$, then the conditional probability of event A, given event B, is defined by,

  $$P\left(A \left| B\right.\right) = \frac{P\left(A \cap B\right)}{P\left(B\right)} \tag{2.4}$$

Note that independence is quite different from disjointness, from which axiom 5 applies. Moreover, if A and B are independent, then $P\left(A \left| B\right.\right) = P\left(A\right)$.

In his aim was to demonstrate that these axiomatization could be used as a rigorous basis for the study of infinite sequences of random variables.

2.2 Probability Axioms

When referred to finite sample spaces, Kolmogorov's axiomatization gets considerably simplified.

In this case, given a sample space Ω and an event A with a probability $P\left(A\right)$, Kolmogorov's axioms are equivalent to the following [6]:

- Axiom 1. for each A, $0 \leq P\left(A\right) \leq 1$.
- Axiom 2. if $A = \oslash$, then $P\left(A\right) = 0$.
- Axiom 3. if $A = \Omega$, then $P\left(A\right) = 1$.
- Axiom 4. $P\left(A \cup B\right) = P\left(A\right) + P\left(B\right)$ iif $A \cap B = \oslash$.

This is really concise (and it can even be stated in a more concise way). But it is not its conciseness what we are interested in. On the contrary, probability is a complex idea and we would have liked Kolmogorov to be more explicit.

Because of this, three standard interpretations of probability, each one in accordance with Kolmogorov's axiomatization, have been closely entangled all along the development of probability theory.

- Frequency. From the frequentest interpretation, $P(A)$ is taken to be the long-run frequency with which A happens in a certain experimental setup or in a certain population. This frequency is something inherent to the experimental setup or the population, and independent of any person's beliefs. (Example dices)

- Belief. This interpretation is close to the former but allows some subjectivity at the time of posting odds for $P(A)$. (Example horse-racing)

- Support. From this interpretation, $P(A)$ is a rational degree of belief to which we should expect A will happen, according to the degree our evidence supports it. (Example database).

Another point of controversy is whether the Kolmogorov's axioms are normative or descriptive, and this may have a special relevance in AI domains.

2.3 Knowledge Discovery Axioms

What is the interpretation of probability from our knowledge acquisition perspective? Does it fit to any of the standard interpretations exposed above?

The answer to this question depends on our ultimate objectives. When we are mining data, we expect to find *useful knowledge*. That is, truthful information that may lead to accurate models of the given domain, in order to make inference or take decisions, or in order to get some understanding about it.

In essence, that means that knowledge is going to be expressed as association patterns between features, eventually bound to a set of conditional probability parameters. Therefore, knowledge is here the answer to two different questions that may, or may not, be combined: knowing *what* and knowing *how*, both of them closely related to the frequencies observed in the sample, and consequently to probability theory.

At first glance, the support interpretation of probability seems to be the closest one. Anyway, we expect to get *whats*, and *whats* relate to facts or essential characteristics of the domain, prior to the support observed in the sample. Hence we approach the frequentest interpretation, seeking for probabilities as something expected to hold in the long run. The belief interpretation should be the furthest. But obviously, if we are willing to have a rational degree of belief based upon evidential support, we are intrinsically facing a subjective question. And here lies the motivation of our work. We pretend to state the right will for this rational degree of belief, so as to avoid any subjectivity.

We may conclude that knowledge seems to be definitely related to probability. Indeed, to all of its classical interpretations. But, we don't subscribe to the whole of this idea.

What are the reasons to believe that knowledge behaves under the Kolmogorov's axioms? Does knowledge about different events combine (addition, product) to give knowledge of them both together? What's the meaning of a knowledge of one? Can we have full certainty in the context of knowledge acquisition? What's the meaning of a knowledge of zero? We can figure out a probability of zero as the impossibility of an event. But we are definitely expressing some knowledge with this assertion.

In short, in a knowledge discovery context the relation $frequencies \rightarrow$ *probabilities*, is not the same as the relation $frequencies \rightarrow$ *knowledge*, or stated in other words *probabilities* $\neq$ *knowledge*.

We already introduced this concept as one of the basic hypothesis in our previous work by the distinction between *structural evidence* and *parametrical evidence*: *structural evidence* refers to the question of whether a relation between two features exist or does not exist, and *parametrical evidence* refer to how this relation operates in case it does exist. This is the meaning of our assertion that *probabilities* $\neq$ *knowledge*.

Then, if we are going to assume that *probabilities* $\neq$ *knowledge*, we may wonder whether probability axiomatization is the most appropriate in this context or could be more conveniently expressed.

In this work we present an approach to knowledge discovery that suggests a somewhat different axiomatization for the behavior of knowledge. We are far from pretending to state an absolute set of axioms for this question, but our work suggests some interesting ideas. Basically, a more appropriate axiomatization for knowledge discovery would be the following slightly modified version of probability axioms:

Given a sample space $\Omega = \{X^p, X^q\}$, for any event $A \in X$ we write $C_A = (|X| - 1) / |X|$ and $U_A = 1/|X|$ as the certainty and uncertainty factors associated to A. Additionally, 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 $A \in \Omega$, $U_A \leq K(A) \leq C_A$

with analogous definitions for conditional knowledge and independence,

- Axiom 2. for any two events $(A, B) \in \Omega$, $U_B \leq K(B|A) \leq C_B$

- Axiom 3. B is independent of A as long as,

$$K(B|A) \leq K(B) \tag{2.5}$$

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

- 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;
- we are not considering any particular algebraic structure for knowledge about disjoint events, therefore no notion of countable additivity is necessary;
- in case a relation of dependence certainly exists, knowledge about the consequent does not depend on knowledge about the antecedent, it depends only on the knowledge conveyed by the pattern;
- 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 $K\left(X^q \left| X^p \right.\right) \geq K\left(X^q\right)$.

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.

2.4 The quality of knowledge

We have stated that we are concerned with *useful knowledge* in the form of patterns or rules, as expressions of *knowledge of a consequent given that an antecedent is known.* Also, we have stated that in the context of data mining, this is inherently related to uncertainty.

From information theory, *uncertainty* is defined from the point of view of a set of possibilities, from which one is expected to be selected. This is called a *scheme of choice* and is denoted by $S = \{e_1, e_2, ..., e_m\}$, [1]. Uncertainty refers to the fact of not knowing which one is going to be selected. Intuitively, the larger the cardinality $|S|$ of S, the larger the uncertainty. Anyway, due to some information based considerations, uncertainty is considered to be much lower, and given by the entropy of S, $H\left(S\right) = log\left(|S|\right)$. This makes perfect sense in this context and sets the basis for a sound theory about information, from which many measures are defined, [7], [4].

From our perspective, we focus *uncertainty* from the idea of discretizing a continuous feature. In this context *uncertainty* takes quite a different sense: a discretization into k intervals renders an uncertainty of $1/k$, hence, the higher the number of intervals, the lower the uncertainty. This is quite opposite to the information theory point of view, though making perfect sense too.

Such a difference is not amazing at all. One thing is to be concerned about *what will be* or *what will happen* or *what's the reason for*, in other words, to make up and down reasoning on some already known model. A different one is to be concerned about *what it is*, that is, to discover the model itself. In each case, uncertainty, and therefore knowledge, are as defined.

Then, being our concept of uncertainty that of a relative lower quality of information, we consequently have to consider different *qualities* of knowledge, being of higher quality that knowledge given by more accurate discretization. And pushing further this idea, we extend this perception of knowledge, not only to continuous, but to any kind of multinomial features.

Some good approaches already exist which deal directly with continuous features. Then, following the previous reasoning, this one should be the right path to the highest quality of knowledge, and one may question the motivations for discretization approaches. But this is not more than a new form of the *Ockham's razor* dilemma: considering and inferring continuous values in an inherently uncertain context does not seem to have any logic, and furthermore, a good discretization of features should be regarded as the first step against variance of models.

Chapter 3

Geometry

3.1 Deviation from minimum information

We have recently introduced an approach to knowledge discovery based on a measure of deviation with respect to minimum information distributions, namely the *perfect marginal distribution* (*pmd*) and the *null conditional distribution* (*ncd*). In all the following let's refer to [2], [3].

Briefly reviewed, given two features X^p with $|X^p| = r$, and X^q with $|X^q| = s$, a reasonable expression of the knowledge conveyed by the pattern $X^p \to X^q$ is the deviation of their conditional distribution $(X^q\,|X^p)$ with respect to the *ncd*, given by,

$$\Delta\left(x_j^q\,|x_i^p\right) = \left(\frac{x_{i,j}^{p,q} - U_s}{U_s}\right)^2 \quad ; \quad 0 \leq\ x_{i,j}^{p,q} \leq\ U_s$$

$$\Delta\left(x_j^q\,|x_i^p\right) = \left(\frac{x_{i,j}^{p,q} - U_s}{C_s}\right)^2 \quad ; \quad U_s \leq\ x_{i,j}^{p,q} \leq 1$$

where, $U_s = 1/s$ is the *uncertainty factor*, $C_s = (s-1)/s$ is the *certainty factor*, and $x_{i,j}^{p,q}$ are the relative conditional frequencies, given by,

$$x_{i,j}^{p,q} = \frac{n_{ij}^{pq}}{n_i^p}$$

We also exposed the intuition that an analogy could be established between marginal and conditional distributions, based on the idea of considering the pattern $\oslash \to X^q$, or better said, $D \to X^q$, being D the sample data. Following the analogy, 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,

$$\Delta\left(x_j^q\right) \equiv \Delta\left(x_j^q \,|D\right) = \left(\frac{x_j^q - U_s}{U_s}\right)^2 \quad ; \quad 0 \leq\ x_j^q \leq\ U_s$$

$$\Delta\left(x_j^q\right) \equiv \Delta\left(x_j^q \,|D\right) = \left(\frac{x_j^q - U_s}{C_s}\right)^2 \quad ; \quad U_s \leq\ x_j^q \leq 1$$

where the x_j^q are the marginal frequencies. [1]

From these measures of deviation we derived the concepts of *coherence* and *presence*, as measures of *reliability* and *representativity* of the sample.

Finally, as a measure of *amount of knowledge conveyed* by a rule, (or *strength of implication* as the closest concept commonly used in the literature), we combined these two concepts in a notion of *utility*, in which both forms of knowledge are assumed to be clashing.

By clashing, we mean that marginal knowledge about a feature is prior to its relation with any antecedent, and it should be wiped away from the measured conditional knowledge. This turned out to be, in essence, how the bias/variance dilemma takes form in our context.

But, what is what we are really measuring? A simple geometric interpretation brings some light to this question.

3.2 Graphical representation of knowledge

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 shown in fig.3.1

[1] Notation may be here somewhat misleading because x_j^q refers either to the attribute-value pair itself as well as to its marginal frequency.

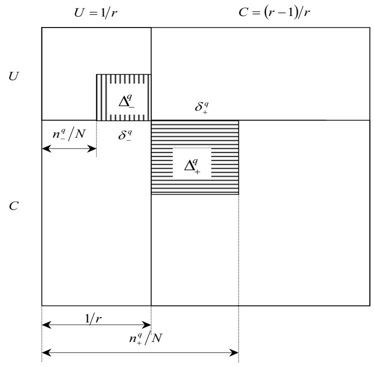

Figure 3.1: Graphical representation of knowledge

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.

Let's note also, that the normalized version of the square deviations, is not more then these areas relative to the maximum knowledge that can be achieved at each side of the point of minimum information,

$$\Delta_- = \frac{\delta_-^2}{U^2}$$
$$\Delta_+ = \frac{\delta_+^2}{C^2}$$

A further illustration of how deviations and square deviations are related in our graphical representation of knowledge is given in fig.3.2.

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,

$$\sum_j^s \delta_j^q = \sum_j^s \left(\frac{n_j^q}{N} - \frac{1}{s} \right) = \frac{1}{N} \sum_j^s n_j^q - 1 = 0 \tag{3.1}$$

that is,

$$\sum_+ \delta_+^q = \delta_3^q = \delta_1^q + \delta_2^q = \sum_- \delta_-^q \; .$$

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

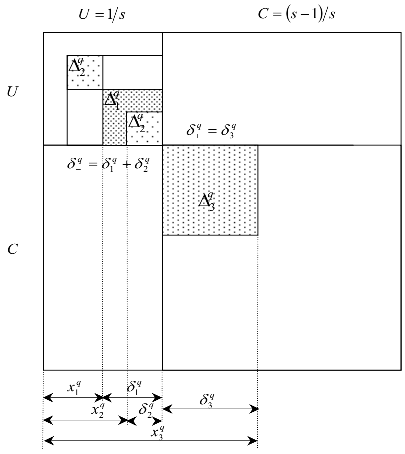

Figure 3.2: Knowledge representation for X^q with $s = 3$

$$\sum_j^s \Delta_j^q = \sum_j^s \left(\delta_j^q\right)^2 = \sum_j^s \left(\frac{n_j^q}{N} - \frac{1}{s}\right)^2 = \sum_j^s \left(\frac{n_j^q}{N}\right)^2 - U_s \tag{3.2}$$

Hence, when we have maximum deviation, *i.e.* $\sum_j^s \Delta_j^q = 1 - U_s$, we still have a lack of knowledge amounting U_s, and what we get is the shadowed areas shown in figure 3.3 for different values of s.

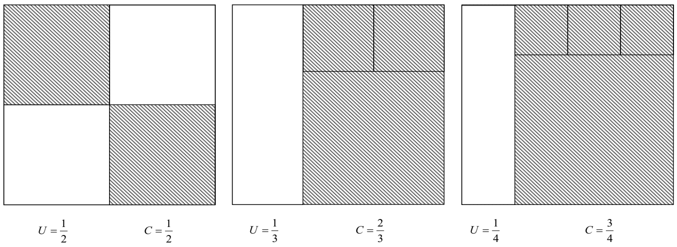

Figure 3.3: Areas of certainty and uncertainty for $s = 2, 3, 4$

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.

An immediate pretty useful conclusion can be derived from this graphical representation. As long as we are uniquely interested on $\sum_j^s \Delta_j^q$, analogous results (though and asymmetrical curve) can be obtained by symmetrical normalization of square deviations at both sides of U with,

$$\Delta\left(x\right)=\frac{\left(x-U\right)^2}{C} \tag{3.3}$$

This is really advantageous at the time of handling these expressions in further definitions or properties, and constitutes a real improvement with respect to our previous approach.

Anyway, in all the following, it is going to be more useful to consider the not normalized areas, so, except when explicitly noted, let's consider the Δ's just as,

$$\Delta_j^q=\left(\delta_j^q\right)^2=\left(x_j^q-U_s\right)^2 \tag{3.4}$$

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?

Let's refer 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,

$$K\left(X^q\right)=U_s+\sum_j^s\left(\delta_j^q\right)^2 \tag{3.5}$$

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. Note that we are considering here not normalized areas.

Still more important is that, this simple expression holds an interesting property related to the cardinality scaling of knowledge, as an inherent effect to discretization of continuous features.

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

Finally, following the same reasoning for each event x_j^q, we would have,

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

and it is worth noting the slight but significant difference that appears, at this point, between knowledge and probabilities (as estimated from raw frequencies), that is,

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

Now, let's refer again to our axiomatic approach,

- The maximum knowledge we can have is that given by the certainty factor.

Again, the most simple expression of this idea is,

$$K\left(X^q\right) = C_s\left(U_s + \sum_j^s \Delta_j^q\right)$$

where the Δ's should now be referred to the certainty factor, that is normalized with respect to C_s,

$$\Delta_j^q = \frac{\left(\delta_j^q\right)^2}{C_s} = \frac{\left(x_j^q - U_s\right)^2}{C_s}$$

with an analogous expression of knowledge for each event x_j^q as,

$$K\left(x_j^q\right) = C_s\left(U_s + \Delta_j^q\right) \tag{3.6}$$

Its minimum and bounder values are given by,

- for $x_j^q = 0 \rightarrow \Delta_j^q = \frac{U_s^2}{C_s}$, thus
$$K\left(x_j^q\right) = C_s\left(U_s + \frac{U_s^2}{C_s}\right) = C_sU_s + U_s^2 = U_s\left(C_s + U_s\right) = U_s$$
- for $x_j^q = U_s \rightarrow \Delta_j^q = 0$, thus $K\left(x_j^q\right) = C_s\left(U_s + 0\right) = C_sU_s$
- for $x_j^q = 1 \rightarrow \Delta_j^q = C_s$, thus $K\left(x_j^q\right) = C_s\left(U_s + C_s\right) = C_s$

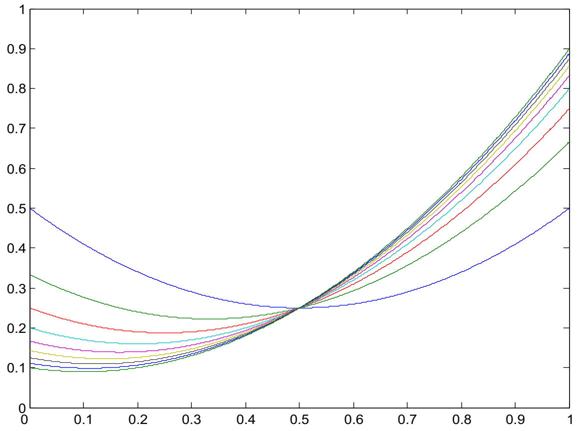

Figure 3.4: Knowledge for $2 \leq s \leq 10$

Finally, in order to get normalized values of knowledge, we should normalize again this expression with respect to its global maximum for all events x_j^q, given by,

$$K\left(X^q\right)_{max} = (s-1)\; U_s + C_s = 2\, C_s$$

what finally renders eq. 3.6 as,

$$K\left(x_j^q\right) = \frac{1}{2}\left(U_s + \Delta_j^q\right) \tag{3.7}$$

In fig. 3.4, we show the depiction of eq. 3.6. This graphical representation is certainly illustrative. It suggests a reference point given by the value of minimum information with cardinality 2, and given a relative frequency of $x_j^q = 1/2$, knowledge is the same, be what it be the cardinality. This looks as a new perspective of our scaling property, once the measures referred to C_s. We show this in fig. 3.5.

Then, from this reference point, and given any deviation ϵ in one or the other sense, we have that for any cardinality s,

- while $U_s \leq x_j^q \leq U_2$, knowledge is higher with cardinality 2, and the lower the cardinality, the lower the knowledge,
- if $x_j^q \geq U_2$, knowledge is higher with cardinality s, and the higher the cardinality, the higher the knowledge.

In other words, this means that whenever there is good information, we would like it to be expressed with the highest possible accuracy, and

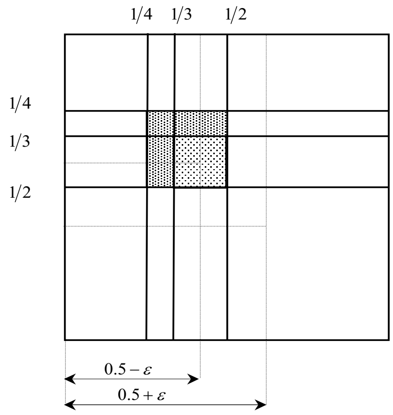

Figure 3.5: $K\left(x_j^q\right)$ for $x_j^q = 1/2$

whenever information is blurred, it is wiser to be more conservative and express it with lower accuracy.

Finally, let's just note that the same reasoning given above applies more appropriately to the conditional distribution $(X^q\,|X^p)$. In this case we have the following normalized expressions of the knowledge conveyed by:

- the whole pattern $X^p \rightarrow X^q$

$$K\left(X^q\,|X_i^p\right) = \frac{1}{2\,r}\left(U_s + \sum_{i,j}^{r,s} \Delta_{ij}^{pq}\right) \tag{3.8}$$

- the subpattern $X_i^p \rightarrow X^q$

$$K\left(X^q\,|X_i^p\right) = \frac{1}{2\,r}\left(U_s + \sum_{j}^{s} \Delta_{ij}^{pq}\right) \tag{3.9}$$

- or the rule $X_i^p \rightarrow X_j^q$,

$$K\left(X_j^q\,|X_i^p\right) = \frac{1}{2\,r}\left(U_s + \Delta_{ij}^{pq}\right) \tag{3.10}$$

where,

$$\Delta_{ij}^{pq} = \frac{\left(x_{ij}^{pq} - U_s\right)^2}{C_s}$$

3.3 Half dependence

In our axiomatic dissertation we also conclude that the dependence or independence condition between any two features, should be better regarded as a relation of half dependence, with a reasonable decision boundary given by the point where the conditional, and the marginal knowledge, about the consequent, are equal.

Consequently, we may consider a measure about their degree of dependence, defined as the difference between these two expressions, that is,

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

what yields,

$$\sum_{i,j}^{r,s} \frac{1}{2\,r}\left(U_s + \Delta_{ij}^{pq}\right) - \sum_{j}^{s} \frac{1}{2}\left(U_s + \Delta_{j}^{q}\right) = \\ \frac{1}{2}\left(\sum_{i,j}^{r,s} \frac{1}{r}\,\Delta_{ij}^{pq} - \sum_{j}^{s} \Delta_{j}^{q}\right)$$

Let's note that, for $X^q \perp X^p$, we should expect that,

$$\forall\left(x_i^p, x_j^q\right)\ ;\ \frac{n_{ij}^{pq}}{n_i^p} = \frac{n_j^q}{N} \Rightarrow x_{ij}^{pq} = x_j^q$$

and therefore,

$$\sum_{i,j}^{r,s} \Delta_{ij}^{pq} = \sum_{i,j}^{r,s} \frac{\left(x_{ij}^{pq} - U_s\right)^2}{C_s} = r\sum_{j}^{s} \frac{\left(x_j^q - U_s\right)^2}{C_s} = r\sum_{j}^{s} \Delta_j^q$$

hence,

$$K\left(X^q \left| X^p\right.\right) - K\left(X^q\right) = 0$$

So, this difference is zero for stochastical independence, is negative whenever marginal knowledge is greater then conditional knowledge, and is positive otherwise.

Chapter 4

Information theory

In our previous work, based on the intuition that some extra scaling should be given to our measure of deviation, we presented the following general expression,

$$Z\left(x_j^q\right) = k\, Q^{\left(\alpha\, \Delta_j^q\right)}$$

where k is a normalizing factor, Q is a scaling factor, α is a shaping factor and Δ_j^q is the area associated to a deviation δ_j^q

Afterward, assuming that knowledge is indeed related with deviation from minimum information, and hence, Δ_j^q represents the piece of knowledge contributed by x_j^q, we thought that we may directly add these pieces.

$$Z\left(X^q\right) = \sum_j^s Z\left(x_j^q\right) = k\, \sum_j^s Q^{\left(\alpha\, \Delta_j^q\right)}$$

Although with important analogies with this, in this work we have presented a different approach, based upon a certain axiomatization of knowledge discovery and a geometric interpretation of knowledge.

In this chapter, we still present a different approach, which sets a bridge between our previous exposition and some information theory based concepts.

As exposed in 2.4, within the framework of information theory, uncertainty is defined as logarithmic. Among other reasons, this contributes an important ease of mathematical handling of expressions at the time of defining concepts and deriving important properties about information.

In this sense, it is good to adopt this idea, and we may wonder whether a better aternative would be to consider the following general expression,

$$Z\left(X^q\right) = \prod_j^s Z\left(x_j^q\right) = k\, Q_s^{\left(\alpha\, \sum_j^s \Delta_j^q\right)} \tag{4.1}$$

or its analog for conditional distributions,

$$Z\left(X^q\,|X^p\right)=\prod_{i,j}^{r,s} Z\left(x_j^q\,|x_i^p\right)=k\,Q_s^{\left(\alpha\,\sum_{i,j}^{r,s}\Delta_{ij}^{pq}\right)} \tag{4.2}$$

where, still assuming that knowledge is related to the global area, we focus on a volumetric idea of it.[1]

4.1 Presence and Coherence

The most important difference between the two approaches is that the latter allows us to handle the knowledge conveyed by a distribution as a whole.

Thus, relaying again on information theory, we can apply the concept of distance between distributions, as a measure of relative knowledge between distributions. That is, given two different distributions $A\,(X)$ and $B\,(X)$,

$$D\,(A\,\|\,B)=log\left(\frac{Z_A\,(X)}{Z_B\,(X)}\right) \tag{4.3}$$

holds the properties of a directed distance [4], and may be considered as the difference in knowledge conveyed by each, being X either a marginal or a conditional distribution.

This allows us to express our previously defined concepts of *presence* and *coherence* from this perspective:

- given any marginal distribution of X^q, we can redefine our concept of *presence* as a relative distance to the point of maximum information, that is with $\sum_j^s \Delta_j^q = s$,

  $$B^q=log\left(\frac{Z\,(X^q)}{Z_{max.}}\right)=log\left(\frac{k\,bxp\left(\alpha\sum_j^s\Delta_j^q\right)}{k\,bxp\,(\alpha\,s)}\right)$$

  what gives,

  $$B^q=-\alpha\,log\,(Q_s)\left(s-\sum_j^s\Delta_j^q\right) \tag{4.4}$$

  and from $B^q=\sum_j^s b_j^q$, for each x_j^q we have,

  $$b_j^q=-\alpha\,log\,(Q_s)\left(1-\Delta_j^q\right) \tag{4.5}$$

[1] Depending on whether Q_s is greater or less then 1 this expression is going to measure certainty or uncertainty. Based on our previous experience we are going to assume that $Q_s<1$.

- given any conditional distribution $(X^q\,|X_i^p\,)$, we can redefine our concept of *coherence* as the relative distance to the point of minimum information, with $\sum_j^s \Delta_{ij}^{pq} = 0$,

$$C_i^{pq} = -log\left(\frac{Z\,(X^q\,|X_i^p\,)}{Z_{min.}}\right) = -log\left(\frac{k\,bxp\left(\alpha\sum_j^s \Delta_{ij}^{pq}\right)}{k\,bxp\,(\alpha\,0)}\right)$$

what gives,

$$C_i^{pq} = -\alpha\,log\,(Q_s)\left(\sum_j^s \Delta_{ij}^{pq}\right) \tag{4.6}$$

and from $C_i^{pq} = \sum_j^s c_{ij}^{pq}$, for each x_{ij}^{pq} we have,

$$C_{ij}^{pq} = -\alpha\,log\,(Q_s)\,\Delta_{ij}^{pq} \tag{4.7}$$

4.2 The scaling and shaping factors

Right till now, we have given some redefinitions of our concepts, but essentially, we still know nothing about our scaling and shaping factors.

By one side, it is obvious that the shaping and scaling factor have a joint effect.

By the other side, it is easy to realize that the behaviour of knowledge, as given in section 3.2, fits well with the assumption given in eq.4.2, and we can easily identify terms by comparing equations 3.9 and 3.10 with equations 4.6 and 4.7, that is,

$$K\,(X^q\,|X_i^p) = \frac{1}{2\,r}\left(U_s + \sum_j^s \Delta_{ij}^{pq}\right) \quad vs. \;\; C_i^{pq} = -\alpha\,log\,(Q_s)\left(\sum_j^s \Delta_{ij}^{pq}\right)$$

$$K\left(X_j^q\,|X_i^p\right) = \frac{1}{2\,r}\left(U_s + \Delta_{ij}^{pq}\right) \quad vs. \;\; C_{ij}^{pq} = -\alpha\,log\,(Q_s)\,\Delta_{ij}^{pq}$$

From this comparison, and as we pointed out in section 3.2, we may conclude that the geometry of our graphical representation of knowledge, certainly states a right cardinality scaling of knowledge, and the scaling factor is indeed replaced by an offset of knowledge, determined by the uncertainty factor. Herein, the roll of the scaling factor would be merely as a unit of measure.

This is so in information theory, where it is common to take 2 as the logarithmic base, so as to have measures in bits. But we had assumed from the beginning that $Q < 1$. Consequently, and due to the directionality of distances, we have to change some negative signs in our expressions.

These considerations render, as the most simple form of equations 4.1 and 4.2, the following expressions,

- marginal knowledge,

$$Z\left(X^q\right) = 2^{-\frac{U_s}{2}}\, 2^{-\frac{1}{2}\sum_j^s \Delta_j^q}$$

- conditional knowledge,

$$Z\left(X^q\left|X^p\right.\right) = 2^{-\frac{U_s}{2\,r}}\, 2^{-\frac{1}{2\,r}\sum_{i,j}^{r,s} \Delta_{ij}^{pq}}$$

Therefore, the particular expressions of *presence* and *coherence* are given by,

- presence,

$$b\left(X_j^q\right) = \frac{1}{2}\left(1 - \Delta_j^q\right)\ , \quad B\left(X^q\right) = \frac{1}{2}\left(s - \sum_j^s \Delta_j^q\right)$$

- coherence

$$c\left(x_{ij}^{pq}\right) = \frac{1}{2\,r}\left(U_s + \Delta_{ij}^{pq}\right)\ , \quad C\left(X^{pq}\right) = \frac{1}{2\,r}\left(U_s + \sum_{i,j}^{r,s} \Delta_{ij}^{pq}\right)$$

Note that, in the expression of *coherence*, the point of minimum information from which we take relative distances, is the absolute minimum given for $s = \infty$. This is a significative difference with respect to our previous approach. Both expressions are depicted in fig.4.1.

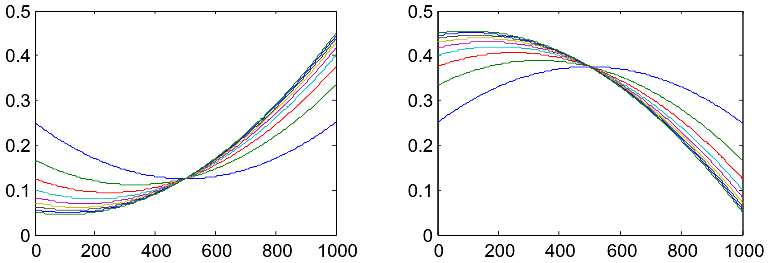

Figure 4.1: Presence and Coherence for $2 \leq s \leq 10$.

4.3 Half dependence revisited

Let's note that, the distance between the conditional and the marginal expressions of knowledge about X^q is,

$$D\left(\left(X^q\left|X^p\right.\right) \parallel X^q\right) = -log\left(\frac{Z\left(X^q\left|X^p\right.\right)}{Z\left(X^q\right)}\right)$$

that is,

$$D\left(\left(X^q\left|X^p\right.\right) \parallel X^q\right) = \frac{1}{2}\left(\frac{1}{r}\sum_{i,j}^{r,s}\Delta_{ij}^{pq} - \sum_{j}^{s}\Delta_j^q\right)$$

just the same expression that we give in section 3.3, where the difference of areas is an expression of the directional degree of dependence between both features.

Chapter 5

Conclusions

We have presented a graphical representation of our measures of deviation to uninformative distributions, supported with some axiomatic considerations about knowledge discovery. Also, setting a bridge to some information theory based concepts, we have given a redefinition of the concepts of *presence* and *coherence*, directly related to their meaning as measures of marginal and conditional knowledge.

This graphical perspective suggests a plausible interpretation of knowledge. Measures of deviation are areas that may be considered as pieces of knowledge, either relative to the absolute knowledge, given by the full square, or relative to the maximum knowledge that can be conveyed by each cardinality. The geometry itself, states the different *quality* of knowledge to each cardinality, as areas of certainty and uncertainty associated to each.

Additionally, this graphical representation of knowledge turns to be a powerful tool. Some interesting cardinality relations are expressed in its geometry. Herein, we have found out that the cardinality scaling of knowledge may be given by an offset of certainty, instead of a direct scaling factor. This is probably the most important difference with respect to our previous work.

In our opinion, the equilibrium expressed by the function of *utility* is now definitely righter then before, but we are not yet sure enough, about being on the *rightest will for believing what we see.* Some different properties are likely to be derived from this geometric relations, and this may contribute with new ideas about the structure, the behavior and the right cardinality scaling of knowledge.

Also, a thorough analysis about how this framework applies to the parametrical evidence functions and to our desiderata of properties, is left for future work. By the moment, nothing has been worked out about this question.

Concluding, we deemed this approach to be specially interesting, not only because it gives better results, as we show in the following examples, but also because it offers a theoretical base which opens a path to further research and development.

5.1 Examples

In our previous work we presented an application of this framework to a synthetic domain with continuous features, with some examples about what we call *domain sensitive discretization.*

Let's remember that, being all three features continuous, the challenge was to find the right discretization of each one, so as to identify the five classes, and the conditional relation of dependence that governs the domain.

Following, we present in figures 5.1 and 5.2, the same examples, in order to show a comparison of the empirical results obtained with the previous and the new approach. The new ones are definitely better.

Basically, let's note that for the unbalanced examples, the utility function allows the search algorithm for a better detection of the five classes. For the highly unbalanced one, this is at the cost of a significant over discretization of the explanation variables, with an interval width of about the same marginal frequencies then those of the less represented classes.

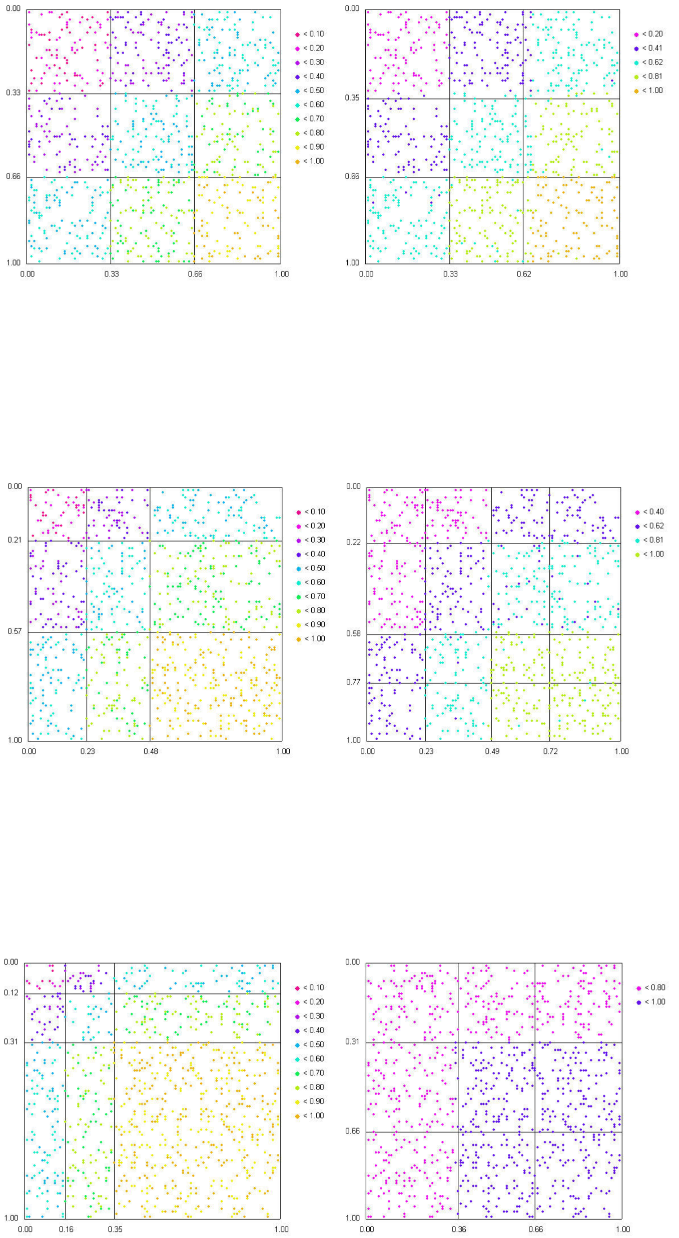

Figure 5.1: Results with the previous approach

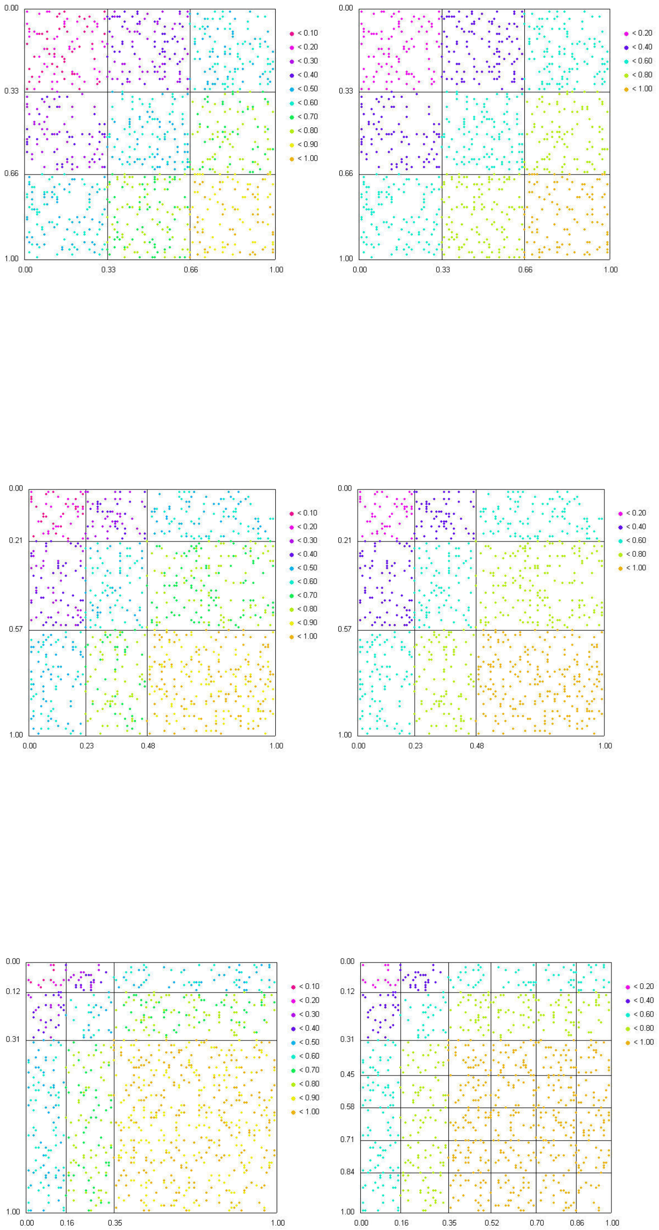

Figure 5.2: Results with the new approach

Bibliography

[1] Bavaud F., Chappelier J.C., Kohlas J. An introduction to Information Theory and Applications. March 2005.

[2] 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.

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

[4] Kullback S. Information theory and statistics. Dover Publications Inc., New York, 1997. (Unabridged republication of John Wiley & Sons, New York, 1959.)

[5] Kolmogorv A.N. Foundations of the theory of probability. Edited by N.Morrison, Chelsea Publishing Company, New York, 1956.

[6] Shafer G.. What is probability? In D.C. Hoaglin & D.S. Moore (Ed.), *Perspectives on Contemporary Statistics.* MAA Notes 21, pp.93-105, Washington, DC:Mathematical Association of America, 1992

[7] Shanon C.E. A Mathematical Theory of Communication. *The Bell System Technical Journal*, vol.27, pp.379-423,623-656, July, October, 1948.

Appendix A

A.1 Cardinality scaling of knowledge

Let's think about a discretization in $(s+1)$ intervals such that one of the categories is empty, and the remaining s categories have each one the same marginal frequencies. This is shown in fig. A.1.

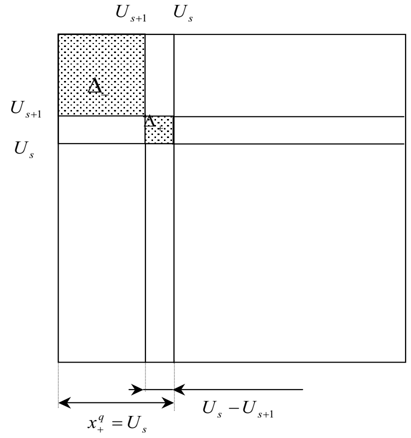

Figure A.1: Cardinality scaling of knowledge

In such situation, we have that for each non empty category, let's generically denote them as x_+^q, the marginal frequencies are, obviously, just the uncertainty factor for cardinality s,

$$n_+^q = \frac{N}{s} \rightarrow x_+^q = \frac{n_+^q}{N} = U_s$$

with an associated deviation of,

$$\delta_+^q = x_+^q - U_{s+1} = U_s - U_{s+1} = U_s \, U_{s+1} \tag{A.1}$$

Therefore, the total square deviation of this marginal distribution is,

$$\sum_{j}^{s+1} \left(\delta_j^q\right)^2 = \Delta_- + s\,\Delta_+$$

$$= U_{s+1}^2 + s\,U_s^2\,U_{s+1}^2$$

$$= U_{s+1}^2\;(1+U_s)$$

$$= \left(\frac{1}{s+1}\right)^2 \left(\frac{s+1}{s}\right)$$

$$= \frac{1}{(s+1)}\,\frac{1}{s}$$

$$= U_{s+1}\,U_s$$

But it is obvious that, in such situation, we should consider the real cardinality as being s, and the feature as being in *pmd*, therefore, our real knowledge about it should be the minimum for s, that is U_s.

This is exactly the scaling that equation 3.5 expresses. For such a case, it gives a total knowledge of, (apply eq. A.1),

$$K\left(X^q\right) = U_{s+1} + \sum_{j}^{s+1} \left(\delta_j^q\right)^2 = U_{s+1} + U_{s+1}\,U_s = U_s$$

A.1.1 Non-subsequent cardinalities

This holds generically for any two cardinalities s and r, being $s > r$. Let's consider a feature of cardinality s, with $(s-r)$ empty categories and the remaining r categories with equal frequencies, given by $x_+^q = U_r$, and an associated deviation of,

$$\delta_+^q = x_+^q - U_s = U_r - U_s$$

In general, and being equation A.1 just a particular case of it, this deviation between any two cardinalities, is given by,

$$\delta_+^q = U_r - U_s = (s-r)\;U_s\,U_r \tag{A.2}$$

Therefore, the total square deviation of this marginal distribution is,

$$\sum_{j}^{s} \left(\delta_j^q\right)^2 = (s-r)\,\Delta_- + r\,\Delta_+$$

$$= (s-r)\,U_s^2 + r\,(s-r)^2\,U_s^2\,U_r^2$$

$$= (s-r)\;U_s^2\;(1 + (s-r)\;U_r)$$

$$= (s-r)\;U_s^2\,s\,U_r$$

$$= (s-r)\,U_s\,U_r \tag{A.3}$$

and the total knowledge conveyed is,

$$K\left(X^q\right) = U_s + \sum_{j}^{s} \left(\delta_j^q\right)^2 = U_s + (s-r)\,U_s\,U_r = U_r \tag{A.4}$$

It is worth mentioning the following relations with respect to the difference in uncertainty between both distributions,

$$\delta_+^q = U_r - U_s = (s-r)\,U_s\,U_r = \sum_{j}^{s} \left(\delta_j^q\right)^2 \tag{A.5}$$

A.1.2 Arbitrary distributions

Now, let's see that it also holds for any arbitrary distribution with one or more empty categories. Such situation is shown in fig.A.2.

As in the previous case, the real cardinality we should consider is r, and the knowledge conveyed by this distribution would be that given by the general expression,

$$K\left(X^{qr}\right) = U_r + \sum_{j}^{r} \left(\delta_j^{qr}\right)^2 \tag{A.6}$$

With respect to cardinality s, the knowledge conveyed is,

$$K\left(X^{qs}\right) = U_s + \sum_{j}^{s} \left(\delta_j^{qs}\right)^2 = U_s + \sum_{j}^{s} \left(\delta_j^{qs}\right)^2$$

where for the $(s-r)$ empty categories, let's generically denote them as x_z^q, the square deviation is U_s^2 for each one, that is,

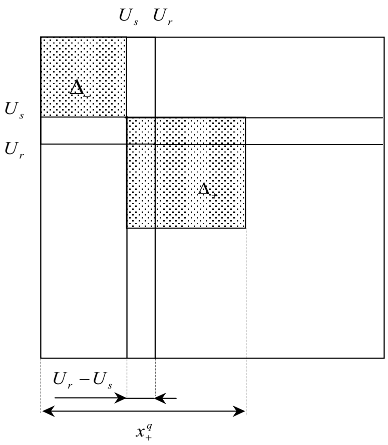

Figure A.2: Cardinality scaling of knowledge

$$K\left(X^{qs}\right) = U_s + (s-r)\; U_s^2 + \sum_{j \neq z} \left(\delta_j^{qs}\right)^2 \tag{A.7}$$

In this expression, the square deviation of the remaining non empty categories is,

$$\sum_{j \neq z} \left(\delta_j^{qs}\right)^2 = \sum_{j \neq z} \left(\delta_j^{qr} + (U_r - U_s)\right)^2$$

$$= \sum_{j}^{r} \left(\delta_j^{qr}\right)^2 + 2\; (U_r - U_s) \sum_{j}^{r} \left(\delta_j^{qr}\right) + \sum_{j}^{r} (U_r - U_s)^2$$

where by equations 3.1 and A.2,

$$\sum_{j \neq z} \left(\delta_j^{qs}\right)^2 = \sum_{j}^{r} \left(\delta_j^{qr}\right)^2 + r\; (s-r)^2\; U_s^2\, U_r^2 \tag{A.8}$$

Then, changing eq. A.8 into eq. A.7,

$$K\left(X^{qs}\right) = U_s + (s-r)\; U_s^2 + r\; (s-r)^2\; U_s^2\, U_r^2 + \sum_{j}^{r} \left(\delta_j^{qr}\right)^2$$

Combining now with eq. A.3 gives,

$$K\left(X^{qs}\right) = U_s + (s-r)\; U_s\, U_r + \sum_{j}^{r} \left(\delta_j^{qr}\right)^2$$

And finally, combining with equations A.4 and A.6, we have,

$$K\left(X^{qs}\right) = U_r + \sum_{j}^{r} \left(\delta_j^{qr}\right)^2 = K\left(X^{qr}\right)$$