The Right Will for Seeing and Believing

Joan Garriga
Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

June 30, 2009

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. Furthermore different interests are met at each task at hand. 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 objectivity. The aim of this work is to ponder an approach to objective KDDB, based on a feature cardinality driven distance measure to uninformative distributions. From this perspective, we propose alternative measures of reliability and representativity combined in a notion of *utility.* The properties and 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 patterns are revealed. Also, optimal cardinalities and optimal subsets of ascendants can be found for any feature, according to a natural bias of the measure toward the MDL principle. As a conclusion, it appears we can assertively capture knowledge from this approach. This may be useful for other knowledge discovery tasks.

Contents

Chapter 1

Introduction

Machine learning, data mining, knowledge discovery or many other related labels, refer all of them to slightly different attitudes facing similar problems which, in essence, are sharing a common conceptual basis: the aim is to mimic such a primary human ability as it is learning from observation. In other words, the goal is to get that kind of knowledge which stems basically from experience and can hardly be expressed as formal theories or mathematical equations.

From the AI community, it is not amazing at all to have such an upright purpose given that this learning framework is specially well suited for computing machines. If machines are good at something that is its ability to *observe* greats amounts of data in a very short time. The same experience achieved by humans after years of observations, can easily be achieved and even improved, by machines, in just a few seconds, whenever these observations are adequately provided. Anyway, as we will show, this question has plenty of loose ends and the learning matter appears itself as a really challenging question.

1.1 Domain representation

If human experience stems from repeatedly observing a fact under different circumstances, adequately providing these observations to machines is bound to two basic ideas: accurately figuring out which are those circumstances that essentially govern the observed fact, here referred to as a *domain*, and optimally decide how to represent such circumstances, usually by means of *features*, taking numerical or categorical values from predefined sets.

This leads us to a first broad division within knowledge discovery, between what is referred to as *transactional* domains and *relational* domains. In *transactional* domains, the features are called *items* and the observations, called *transactions*, are given as *sets of items*. All items are binary, and in each set of items a value of one/zero indicates the presence/absence of that item in the transaction. On the other hand, in *relational* domains, the features are called *attributes*, which are multinomial variables, and each observation is an instantiation of all of them, (at first), in what is called *attribute-value* pairs.

The scope of these different ways of domain representation might be considered somewhat overlapped but there will usually be a better suited one for each task at hand. Therefore, another important reason for this division concerns to the different algorithmic approaches developed for, and applied to, each, and the kind of results that each one can offer. So, the intrinsic characteristics of the domain, along with the kind of results expected, is what will finally determine the better way of representing it.

1.2 The question of learning

The main idea behind any knowledge discovery approach is to find traces of relationship among features that may suggest the cause-effect relations governing the real domain represented. Consequently, we should have available some way of numerically measuring those relationships. Indeed, as we will show, many of them have been developed. Such a comprehensive literature exists about the so called *interestingness measures* that this topic is becoming, day by day, a new discipline by itself.

So, at that point one may wonder, why so many of them exist when there is a single thing to be measured? The answer to this question is definitely not easy and has two loose ends: (i) the first one is *what* should particularly be measured, and refers to a matter of interest, therefore subjective to each particular task at hand, and (ii) the second one is *how* should it exactly be measured, so related to a matter of assertiveness and therefore, I would not dare to say strictly objective, but not so subjective.

In the next chapter we analyze in depth both issues, though it is worth forwarding that none of them is definitely stated and no approaches exist that give optimal results for all types of knowledge discovery related problems.

These relations between features are commonly called *rules*. Rule discovery has received significant research attention because rules are among the most expressive and human understandable representation of knowledge. To state it in a few words, a rule is an implication of the form $a \rightarrow b$, where a and b are either sets of items or attribute-value pairs, bound to a certain degree of reliability, derivable from the fact of observing some degree of coocurrence of a and b.

Another powerful and human understandable representation of knowledge is *graphical models*. In essence, a graphical model is a visual representation of the set of rules that govern the behaviour of the domain, in which each feature is explained as a consequent in a subset, (possibly empty for a marginal independent feature), of these rules.

From an algorithmic perspective, the first broad division can be given in terms of *association rule mining*, and *classification*, and while being true that some fitting exists between this two frameworks and both kinds of domain representation, again we run into some overlapping. We would rather say that this division is more related to the resulting set of rules that one may expect:

1. *classification*, or *machine learning*, is an entire discipline by itself, including *neural nets*, *bayesian nets*, *induction trees*, among other supervised learning methods, initially addressed to relational domains. Anyway, from the machine learning community, some heuristic algorithms for rule discovery have been developed, such as C4.5 rules [29], CN2 [8] or RIPPER [9]. These algorithms focus on classification accuracy and usually return a small set of rules. But their heuristic method does not guarantee the discovery of the best quality rules.

2. in contrast, *association rule mining*, is an unsupervised paradigm that produces a complete rule set, what is more desirable whenever it is computationally feasible. It was initially formulated by [1] for transactional domains and has evolved to a general purpose rule discovery scheme with wide applications, one of them being *class association rules*, fully stepping into classification domains [23]. However, association rule discovery is bound to some user prefixed support and confidence constraints, and usually produces too many rules, being quite inefficient when the minimum support is low. This has account for a lot of research, leading recently to optimal rule set pruning strategies, bound to the development of *interestingness measures* and the study of their analytic and algorithmic properties [6].

Placing our work with respect to these frameworks, it is worth forwarding that we will rather remain in this no man's land of the *class association rules*, but with the difference that any domain feature can take here the roll of a class attribute. We are mining for rules, but these are just only the basic pieces of a more general expectation which is the full pattern that explains each feature. This difference goes even further, in the sense that knowledge discovery is viewed from a *domain sensitive* (nor supervised, neither unsupervised) perspective. Thus, all possible dependencies among features, including conditional dependence relations, can simultaneously be taken into account.

Therefore, though some ideas are exposed with a certain flavour to a particular framework, the most of our work carries through, as a general discovery tool, suitable to both.

1.3 The problems with data

Regardless of domain representation and algorithmic approaches, further problems may appear at learning time, due to some inherent questions related to data driven domain representations.

Unfortunately, it is more usual than desired that the set of observations presents some lack of information: data can be incomplete (*missing values*) and/or unbalanced (skewed marginal frequencies). This problem seriously affects heuristic approaches to relational domains, specially when a multinomial sampling scheme is considered. In this case, the degree of incompleteness may bias its behaviour, unless some actions are taken in order to manage this lack of information. Other undesired effects of this problems in data are overfitting and accuracy degradation.

In association rule mining none of these is considered a problem by definition. From our opinion, unbalanced data may lead to not sufficiently contrasted information, so we step aside from this consideration. Beyond single rules, looking ahead for assertive explanation patterns, unbalance of data becomes a crucial matter. We widely object this all along this work.

Another common problem with data is the presence of continues features or categorical ones with a large set of possible categories. Association rule mining, and most of the classification approaches, can deal only with discrete values, and large cardinalities may lead to prohibitive computational cost. That means that continuous attributes have to be discretized and large categorical attributes have to be clustered. Many methods exist in order to do this and the learning process will be clearly affected by the resulting partitions given by this discretization or clustering procedures. Most often, discretization is done independently for each feature. However, since they may be not independent, one should consider the possibility of simultaneous discretization. We will show how we can deal with this matter from our framework.

Still another question is the relevance of each feature as an information source about the represented domain. In some cases, too many features are available and it is a good approach to get rid of the most irrelevant ones. In other cases, some of them may even be self-defeating, leading to cheating results or lower accurate models. This issue is known as *feature subset selection*. We also will show how the proposed framework can give us some orientation about how to perform this task.

1.4 Notation and terminology

Because of its higher level of generality and its direct translation to association rule mining, we are going to use a relational framework notation.

Let's consider a domain characterized by a set of m multinomial features $X = \left\{X^1, X^2,..., X^m\right\}$ and a set $\{D\}$ of N examples over these features. Let's consider two any features, or attributes, 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 $crd\,(X^p) = r$ and $crd\,(X^q) = s$, 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 $\{D\}$.

Translating to association rule mining, features X^p and X^q are items with cardinalities $r = 2$ and $s = 2$, with possible outcomes $x_1^p = x_1^q = present$ and $x_2^p = x_2^q = absent$.

In order to avoid any misleading interpretation, let's point out that cardinality is sometimes referred, in association rule mining, to the marginal frequency of an item or set of items. It is important to keep this difference in mind, because in our context, the cardinality becomes a fundamental property of the features.

Additionally, and for the purpose of clarity in our exposition, 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 the relation $x_i^p \rightarrow X^q$, 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).

Again, please keep this in mind because, in association rule mining, a pattern refers to a particular set of items in the antecedent of a rule, what is not our case.

Chapter 2

Measures for knowledge discovery

Once stated that we are interested in numerically measure the degree of relationship between features, we must be aware of the following: the only thing that can objectively be measured is *coocurrence* of itemsets or attribute-value pairs, whatever be the case. Down from here, we can implement this information in multiple formal expressions, in order to enhance different aspects of interest, but at that point we are already trespassing the subjective/objective borderline.

2.1 Interestingness criteria

A great amount of different flavoured measures exist, each one associated with some conceptual label indicating the intention lying at the origin of the formal expression. The data mining community refers with the general expression of *interestingness* to the different specific criteria intended to be measured.

A good categorization of interestingness is given in [13], where the authors point up and give detailed descriptions of the following criteria: *conciseness*, *coverage*, *reliability*, *peculiarity*, *diversity*, *novelty*, *surprisingness*, *utility* and *actionability*. As the authors appropriately mention, some of these criteria are somewhat correlated, rather than independent of one another, so it is not an excluding classification. It is more of a clear symptomatic expression about the difficulties associated to the fact of turning a simple coocurrence based information into a conceptual measure of interestingness.

Still from [13], a further classification of these nine criteria is given, splitting them into *objective* (raw data based), *subjective* (data and domain knowledge based) or *semantics-based* considerations, and again some overlapping is unavoidable.

A handful of some of the most commonly used interestingness measures is presented in fig. 2.1 extracted from the same survey given in [13].

Let's forward that, from the most objective considerations, our interest overlaps, mainly, the notions of conciseness, coverage, reliability, utility and actionability:

1. *conciseness*, because the result of a mining process should be strictly restricted to the most relevant information. Much research has been conducted upon this topic, leading to *minimum description length* approaches or pruning approaches such as *monotonicity* [27], *confidence invariance* [3] and more recently *optimonotonicity* [6]. We will show later how our proposal deals with this question from a quite different approach for relational domains. And regarding to transactional domains, some work is ongoing upon the algorithmic properties of the measure, in order to establish pruning strategies.

2. *coverage*, because we believe that assertive knowledge involves any information that is properly faced up to its counterpart. Again, our proposal deals differently with this topic, defining a measure focused in this notion of *contrastability* instead of the common notion of *generality.* This can be useful in order to discover poorly supported but well contrasted information that might be overlooked by other knowledge discovery mechanisms, as well as filtering well supported but not sufficiently contrasted information. Furthermore, in the root of this difference lies a quiet different perspective: within the association rule mining framework, coverage is regarded from the generated rule set with respect to the examples given in the data set, and the generation of an optimal rule set is based upon rule excluding and example uncovering considerations. From our perspective, coverage is regarded from the examples given in the data set with respect to the real domain, thus considering the very likely possibility that not all the information about the domain will be equally represented in the data set, specially in the case of skewed or small and sparse data.

3. *reliability*, because our interest goes obviously to truthful information, that may lead to highly accurate inference models of the given domain. Once more, a lot of proposals have been given from probability, statistics and information retrieval, to measure the reliability of discovered knowledge, relaying basically on confidence or independence considerations. We also contribute here introducing the notion of the quantity or quality of information conveyed by a particular relationship, as a function of the cardinality of the features involved in it.

4. *utility* and *actionability*, as related to the preceding one, because we expect to model the represented domain in order to make inference

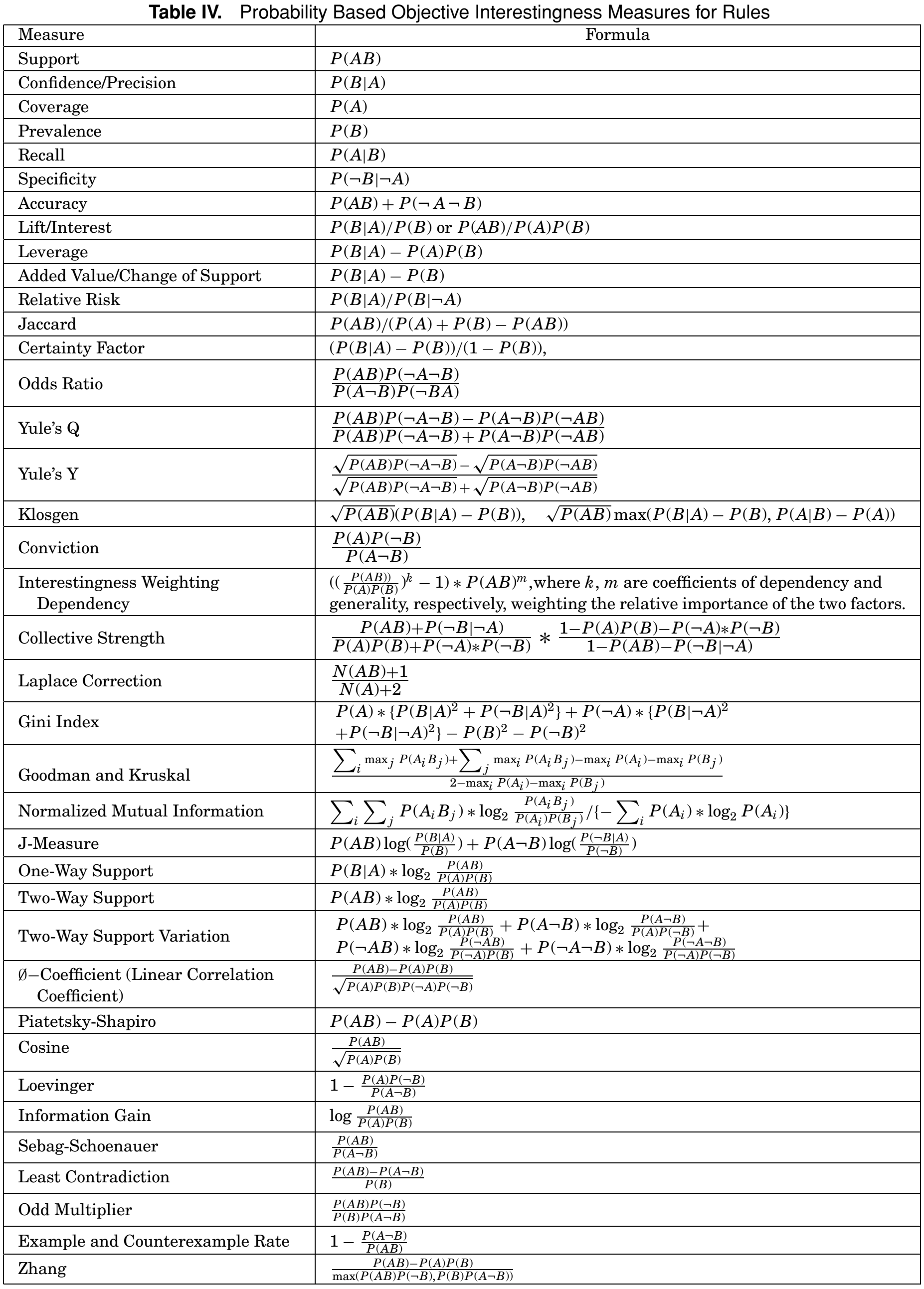
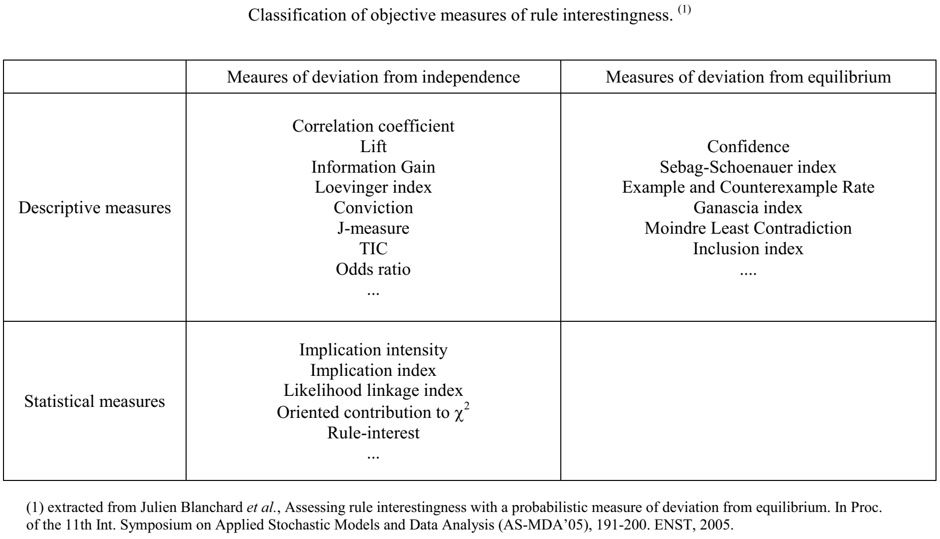


Table IV. Probability Based Objective Interestingness Measures for Rules

Measure	Formula
Support	$P(AB)$
Confidence/Precision	$P(B\mid A)$
Coverage	$P(A)$
Prevalence	$P(B)$
Recall	$P(A\mid B)$
Specificity	$P(\neg B\mid\neg A)$
Accuracy	$P(AB)+P(\neg A\neg B)$
Lift/Interest	$P(B\mid A)/P(B)$ or $P(AB)/P(A)P(B)$
Leverage	$P(B\mid A)-P(A)P(B)$
Added Value/Change of Support	$P(B\mid A)-P(B)$
Relative Risk	$P(B\mid A)/P(B\mid\neg A)$
Jaccard	$P(AB)/(P(A)+P(B)-P(AB))$
Certainty Factor	$(P(B\mid A)-P(B))/(1-P(B))$,
Odds Ratio	$\frac{P(AB)P(\neg A\neg B)}{P(A\neg B)P(\neg BA)}$
Yule's Q	$\frac{P(AB)P(\neg A\neg B)-P(A\neg B)P(\neg AB)}{P(AB)P(\neg A\neg B)+P(A\neg B)P(\neg AB)}$
Yule's Y	$\frac{\sqrt{P(AB)P(\neg A\neg B)}-\sqrt{P(A\neg B)P(\neg AB)}}{\sqrt{P(AB)P(\neg A\neg B)}+\sqrt{P(A\neg B)P(\neg AB)}}$
Klosgen	$\sqrt{P(AB)}(P(B\mid A)-P(B)), \quad \sqrt{P(AB)}\max(P(B\mid A)-P(B), P(A\mid B)-P(A))$
Conviction	$\frac{P(A)P(\neg B)}{P(A\neg B)}$
Interestingness Weighting Dependency	$((\frac{P(AB))}{P(A)P(B)})^k-1)*P(AB)^m$,where k, m are coefficients of dependency and generality, respectively, weighting the relative importance of the two factors.
Collective Strength	$\frac{P(AB)+P(\neg B\mid\neg A)}{P(A)P(B)+P(\neg A)*P(\neg B)} * \frac{1-P(A)P(B)-P(\neg A)*P(\neg B)}{1-P(AB)-P(\neg B\mid\neg A)}$
Laplace Correction	$\frac{N(AB)+1}{N(A)+2}$
Gini Index	$P(A)*\{P(B\mid A)^2+P(\neg B\mid A)^2\}+P(\neg A)*\{P(B\mid\neg A)^2$ $+P(\neg B\mid\neg A)^2\}-P(B)^2-P(\neg B)^2$
Goodman and Kruskal	$\frac{\sum_i \max_j P(A_iB_j)+\sum_j \max_i P(A_iB_j)-\max_i P(A_i)-\max_i P(B_j)}{2-\max_i P(A_i)-\max_i P(B_j)}$
Normalized Mutual Information	$\sum_i\sum_j P(A_iB_j)*\log_2\frac{P(A_iB_j)}{P(A_i)P(B_j)}/\{-\sum_i P(A_i)*\log_2 P(A_i)\}$
J-Measure	$P(AB)\log(\frac{P(B\mid A)}{P(B)})+P(A\neg B)\log(\frac{P(\neg B\mid A)}{P(\neg B)})$
One-Way Support	$P(B\mid A)*\log_2\frac{P(AB)}{P(A)P(B)}$
Two-Way Support	$P(AB)*\log_2\frac{P(AB)}{P(A)P(B)}$
Two-Way Support Variation	$P(AB)*\log_2\frac{P(AB)}{P(A)P(B)}+P(A\neg B)*\log_2\frac{P(A\neg B)}{P(A)P(\neg B)}+$ $P(\neg AB)*\log_2\frac{P(\neg AB)}{P(\neg A)P(B)}+P(\neg A\neg B)*\log_2\frac{P(\neg A\neg B)}{P(\neg A)P(\neg B)}$
$\emptyset-$Coefficient (Linear Correlation Coefficient)	$\frac{P(AB)-P(A)P(B)}{\sqrt{P(A)P(B)P(\neg A)P(\neg B)}}$
Piatetsky-Shapiro	$P(AB)-P(A)P(B)$
Cosine	$\frac{P(AB)}{\sqrt{P(A)P(B)}}$
Loevinger	$1-\frac{P(A)P(\neg B)}{P(A\neg B)}$
Information Gain	$\log\frac{P(AB)}{P(A)P(B)}$
Sebag-Schoenauer	$\frac{P(AB)}{P(A\neg B)}$
Least Contradiction	$\frac{P(AB)-P(A\neg B)}{P(B)}$
Odd Multiplier	$\frac{P(AB)P(\neg B)}{P(B)P(A\neg B)}$
Example and Counterexample Rate	$1-\frac{P(A\neg B)}{P(AB)}$
Zhang	$\frac{P(AB)-P(A)P(B)}{\max(P(AB)P(\neg B),P(B)P(A\neg B))}$

Figure 2.1: Some commonly used objective measures of interestingness



Classification of objective measures of rule interestingness. [1]

	Meaures of deviation from independence	Measures of deviation from equilibrium
Descriptive measures	Correlation coefficient Lift Information Gain Loevinger index Conviction J-measure TIC Odds ratio ...	Confidence Sebag-Schoenauer index Example and Counterexample Rate Ganascia index Moindre Least Contradiction Inclusion index ....
Statistical measures	Implication intensity Implication index Likelihood linkage index Oriented contribution to χ^2 Rule-interest ...	

(1) extracted from Julien Blanchard *et al.*, 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 (AS-MDA'05), 191-200. ENST, 2005.

Figure 2.2: Classification of Objective measures

or take decisions, and also figuring out some understanding of its behaviour, specially for relational domains.

2.2 Measures of deviation

A different classification of interestingness measures is that given by [4] and shown in fig. 2.2, with some examples of the most well known measures.

In this work, the authors state that there exist two different but complementary aspects of rule interestingness: deviation from *independence* and deviation from what they call *equilibrium*, that is the situation of maximum uncertainty of the consequent given that the antecedent is known. Furthermore, they distinguish between descriptive measures and statistical measures.

This classification is further revised in [20] where the authors describe a unified probabilistic framework toward a systematic generalization of association rule measures, related to different reference situations, that is, independence, indetermination or a minimum confidence threshold.

To the most of our knowledge, comprehensive comparative studies of interestingness measures have been done only from the point of view of association rule mining, thus the wider scope of our proposal does not exactly fit into them. Anyway, we definitely have a special predilection for this classification point of view, because in our opinion, it clearly hits the nail on its head: being knowledge acquisition our main goal, we are much more interested in measuring the amount of information conveyed by a relationship among features, than stating its dependence or independence.

Of course, deviation from independence is closely related to conveyed information. But, measures based on deviation from independence have two major pitfalls:

1. Regarding specially to statistical measures, they center their main attention in being as accurate as possible at the time of deciding dependence or independence, that is handling Type-I and Type-II errors. Beyond the type of statistic used, the main focus is placed on methods based on the Bonferroni corrections and so on. And aiming at this, they tend to be less sensitive to degrees of half dependence, resulting in a kind of high accurate switching behaviour. In this sense, descriptive measures would be better well suited.

2. But, there's still a second pitfall. Admitting that what we are concerned about is knowledge, and knowledge means, in our case, certainty about a consequent given that we know an antecedent, any measure based on deviation from independence is biased by definition. Simply, because it is pointing to the wrong direction. The ground floor in order to adeptly measure this concept is, not the independence, but the *equilibrium*, or what, for obvious reasons, we prefer to call *minimum information.*

2.3 Some objections

Though being quite certain that different interests entail (in reference to section 1.2, different *whats* and different *hows*, while talking about knowledge as possible cause-effect relations suggested by the sample, there should be no space left to any possible confusion. The flurry of different measures, the comprehensive literature on selecting the right ones for each task at hand [22],[34], as well as the claim against the hype of omnipotent interestingness measures [35], is no more than a symptom that some lacking hovers the data mining scene.

In our opinion, three basic objections are the culprit: (i) the random 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.

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

[1] This example is extracted from [13].

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

Table 2.1: 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?

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 the argument therein is that though we are talking about objective measures, there is a degree of subjectivity in this way of considering evidential support. As pointed out by G.Shafer in his *Mathematical Theory of Evidence*, (1976), a non-100% reliable system giving a 100% probability for event (A) leaves open the question of which is the probability for (A) not happening, which obviously questions its own prediction of (A) happening. As far as we cannot be sure of it, a sample, and consequently any rule taken out of it, should be considered objectively as having less than a 100% reliability. 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 *confidence*?

The Bayesian approach deals naturally with this question but implies the drawback of assessing the prior distributions, which are hardly known most of the time, rendering measures of support meaningless in the best of cases. The so-called *Dempster and Shaffer Theory* of upper and lower bounds for probabilities, is another useful mathematical formalism to deal with this philosophical aspect of probability, though its practical application may be a little cumbersome. The *iterative proportional fitting* algorithm given by Mosteller is yet another alternative attempt to deal with it, though somewhat contrived. As we will show, the proposed measure offers a new point of view, from which we can approach objectivity more easily and naturally.

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 (in principle), but the lesser the prior probability of finding that rule in the dataset.

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

Chapter 3

Measuring objectively

One really objective 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.

Following, we state some intuitive ideas which form the basis of our approach and should hopefully lead us to fulfill these requirements. Afterward, we introduce the basic idea of *what* we intend to measure, and *how* we pretend to measure it.

3.1 Initial hypothesis

On the basis of our approach lie three intuitive assumptions, related to the concepts of *evidence* and *knowledge*, as we are considering them throughout this work.

3.1.1 Evidence quantifiability

The first hypothesis expresses the idea that the evidence conveyed by a sample, with respect to the domain from which it has been drawn, is finite, and therefore measurable. Consequently, the accuracy of any model or set of rules build up from it, has to be directly related to this amount of evidence.

Practically speaking, this means that we should be able to give a numerical reference that allow us to qualify the quality of a sample, as a function of the features considered, the given partitions of any clustered or discretized ones, and the relations considered among them.

3.1.2 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).

3.1.3 Certainty and uncertainty factors

The third hypothesis expresses the idea that different qualities of knowledge exist. If we can think about knowledge as something measurable, then a rule conveys different quantities or qualities of knowledge, as a function of the cardinalities of the features involved. In other words, knowledge is not a plain matter of reliability. It is also directly related with concepts like certainty, accuracy, precision or rigour, all of them obviously related with the cardinality, tending to exactness in the case of continuity.

Discretization is a very illustrative example of this idea: if we are going to discretize a continuous feature, (keep in mind that from our perspective any feature is a sort of class attribute), the information that we get about that attribute is quite different whether the discretization is made in two classes or in whatever greater than two. Thus, the question is how we can express these different degrees of certainty.

The closest notion to certainty, (in this case its opposite, uncertainty), comes from the information theory based on Shanon's work *A Mathematical Theory of Communication* [32]. We know that in binary domains, where information is coded with binary devices, the capacity for coding information increases exponentially with the number of devices. In his work, Shanon stated that the uncertainty entailed by this coding increases logarithmically, as given by the entropy of the coding device.

For instance, for an n bits coding device, the quantity of information that can be coded (number of possible messages in information theory terminology) is 2^n, and in the case of being all of them equally probable, the uncertainty associated to this coding device is,

$$H = -\sum_i p_i log\left(p_i\right) = \, -2^n\,\frac{1}{2^n}\,log\left(\frac{1}{2^n}\right) \, = \, log\left(2^n\right)$$

In this expression, Shanon was considering the probability of fail of each bit in the coding device, so the total uncertainty had to be proportional to the number of bits needed to code the total information, given by $log\left(2^n\right)$. This is the intuitive reason for which he took a logarithm as an expression of uncertainty, analogously to certain formulations of entropy taken from statistical mechanics. This is also the basis for all entropy based measures.

From our opinion, being this theory perfectly suited for binary domains, it should be carefully considered when applied to multinomial domains.

Any feature can be considered as an s-ary coding device, being s its cardinality. In this case, the capacity of this device for coding information is just its cardinality, (s messages, a linear function of it). Being all of them equally probable the former expression would be,

$$H = -\sum_i p_i log\left(p_i\right) = \, -s\,\frac{1}{s}\,log\left(\frac{1}{s}\right) \, = \, log\left(s\right)$$

But, regarding to uncertainty, we should be aware that we are talking about a different concept. Does it make any sense to consider the probability of fail of $log\left(s\right)$ bits in this case? We are not thinking about retrieving any message from a coding device with a possibility of fail. In a multinomial sampling domain, uncertainty refers only to the probability about a particular outcome of a feature. So, in this case, uncertainty decreases as $1/s$.

As opposite to uncertainty, certainty can be thought of as the number of possible outcomes of the feature that can be discarded once we have evidence of its real outcome. That is, $(s-1)$ out of s.

So, concluding this idea, our hypothesis states that the quantity or quality of knowledge associated to a feature as a function of its cardinality, has to be directly related to these certainty, $C \equiv \left(s-1\right)/s$, and uncertainty, $U \equiv 1/s$, factors.

3.2 Reliability

What do we intend to measure? As explained in the previous section, we are interested in *useful* and *actionable* knowledge, hence in the sense of *certainty about a consequent given that and antecedent is known.* This kind of knowledge allows us to make inference, take decisions and get some understanding about the domain in question.

And *how* can we measure it? We also have exposed that a good measure for these criteria should be based on deviation from *equilibrium*, so taking as a reference a state of *minimum information.*

3.2.1 The null conditional distribution (*ncd*)

This leads as to the formal definition of a reference uninformative distribution, given as a conditional distribution from which no information can be grasped.

Definition 1. Two features $(X^p, X^q) \in X$, with $crd\,(X^q) = s$, 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 $n_{ij}^{pq} = n_i^p / s$

This is an ideal situation, possible only among features with equal cardinality, but clearly defines a state of minimum information.

3.2.2 Deviation from the *ncd*

From this definition, it is immediate to measure a deviation of the conditional distribution of $(X^q \,|X^p)$ with respect to the *ncd*, as

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

Furthermore, it would be good to normalize this measure. But this normalization is not symmetric, and leads us to,

$$\Delta\left(X^q \,|X^p\right) = \sum_{i,j} \left(\frac{n_{ij}^{pq} - \frac{n_i^p}{s}}{0 - \frac{n_i^p}{s}} \right)^2 = \sum_{i,j} \left(\frac{\frac{n_{ij}^{pq}}{n_i^p} - \frac{1}{s}}{\frac{1}{s}} \right)^2 \quad ; \quad 0 \leq \frac{n_{ij}^{pq}}{n_i^p} \leq \frac{1}{s}$$

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

This gives us a kind of confirmation about our third hypothesis: in case of $n_{ij}^{pq} = n_i^p$, that is, maximum certainty, the value of deviation is just the certainty factor $(s-1)\,/s$, and in case of maximum uncertainty, the value of deviation is the uncertainty factor $1/s$.

3.2.3 Differences with respect to χ^2

It is notable that at first glance, this expressions may have a certain flavour to the χ^2 statistic:

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij}^{pq} - n_i^p \frac{n_j^q}{N}\right)^2}{n_i^p \frac{n_j^q}{N}} \quad vs. \quad \Delta = \sum_{i,j} \left(\frac{n_{ij}^{pq} - n_i^p \frac{1}{s}}{n_i^p \frac{1}{s}} \right)^2$$

But there are two important differences: (i) the factor $n_i^p \frac{n_j^q}{N}$ in χ^2, is the joint distribution in the case of independence, while the factor $n_i^p \frac{1}{s}$ in Δ, is the joint distribution in the case of minimum information, and (ii) while the χ^2 is properly normalized in order to have a χ^2 distributed statistic, in our measure, normalization is done simply with respect to one, not aiming at building up any statistic, but rather at having a scalar measure of deviation.

In short, we are not seeking a value to perform an hypothesis test in order to decide dependence or independence for a certain single rule. What we are seeking is a measure of deviation in order to have a joined ranked classification of all possible rules extracted from a mining process.

Obviously, as a counterpart, this value will not allow us to decide dependence or independence, but in our case that is not the matter.

3.3 Representativity

A definitely new contribution of this approach is to extend the same formalism to marginal distributions.

The marginal distribution of any feature is, by itself, a source of knowledge, in the sense that, the more biased it is, the more predictable is the outcome of that feature, independently of the existence of any related antecedent. So, in the same way that we talk about reliability of rules, we can talk about reliability of features, what we call the representativity.

3.3.1 The perfect marginal distribution (*pmd*)

The observation of this analogy, leads us to the formal definition of a reference uninformative distribution for marginal distributions.

Definition 2. A feature $X^p \in X$, with $crd\,(X^p) = r$, 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 $n_i^p = N/r$

3.3.2 Deviation from the *pmd*

Analogously to the previous, the marginal distribution deviation of feature X^p with respect to the *pmd* is given by the following expressions,

$$\Delta\left(X^p\right)=\sum_i\left(\frac{n_i^p-\frac{N}{r}}{0-\frac{N}{r}}\right)^2=\sum_i\left(\frac{\frac{n_i^p}{N}-\frac{1}{r}}{\frac{1}{r}}\right)^2\quad;\quad 0\leq\frac{n_i^p}{N}\leq\frac{1}{r}$$

$$\Delta\left(X^p\right)=\sum_i\left(\frac{n_i^p-\frac{N}{r}}{N-\frac{N}{r}}\right)^2=\sum_i\left(\frac{\frac{n_i^p}{N}-\frac{1}{r}}{\frac{r-1}{r}}\right)^2\quad;\quad \frac{1}{r}\leq\frac{n_i^p}{N}\leq 1$$

3.4 The bias/variance dilemma

First of all, let's point out that *pmd* and *ncd* state two perfectly defined points, completely independent of the sample considered.

Second, we should kept in mind that, regardless of the distribution considered, these expressions are measuring exactly the same concept, and in both cases, it can be regarded as knowledge related to the pattern.

Normally, our interest will be on the conditional distribution, because this conveys the useful knowledge. But we must be aware that knowledge conveyed by the marginal distribution acts in a clashing way: it is implicit knowledge that we already had before considering any relationship, so it should be wiped away from our degree of believe on this relationship. That means that while our interest goes to rules as further as possible from the *ncd*, it also goes to rules involving features as closer as possible to the *pmd*.

In other words, deviation from the *ncd* and deviation from the *pmd* express opposite measures of knowledge. We show this in fig.3.1 where we have depicted the former versus the opposite of the latter.

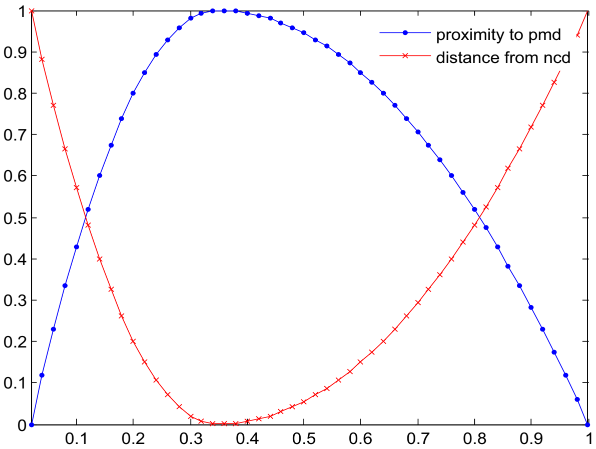


Figure 3.1: Distance from *ncd* and Proximity to *pmd* for s=3

This depiction expresses a kind of balance that exist between this two concepts, given by the convexity of one and the concavity of the other. Recalling the example of discretization, in order to get a partition showing correctly the evidence present in the sample, it is necessary to find a point of equilibrium between reliability and representativity, the optimization of which may become clashing. Optimizing only reliability, may lead to a single interval, a senseless situation expressing zero information. Optimizing only representativity, will most of the times distort reliability.

We are coming across here with our third requirement expressed at the beginning of this chapter. A fair balance must exist between this two concepts and therefore, the importance of being both measured exactly in the same way. As we will further expose in the next section, this is not more then the way the *bias/variance dilemma* takes form within our approach. The main contribution of our work, is to compose this balance in a single measure. This is what we express in our title, (in a literary allowance), as *the right will for seeing and believing.*

At the same time, while composing this balance in a single measure, we are tackling the first requirement. The reliability of the sample is directly related with the degree of coverage of each feature. Ideally, if all features in a sample were in *pmd*, all rule's prior probability would be maximally equilibrated and only in this case, a comparison of their reliabilities would be really objective. Including representativity (or lack of it) in our measure, we are taking into account the reliability of the sample.

In other words, we pretend to state a ground zero at an hypothetical sampling scheme with fixed equal rows and columns. This is obviously an ideal sampling scheme, not likely to have in practice. But no one would argue that it is the most powerful statistical situation in order to get the most reliable conclusions. So, our intention is to state it as the absolute reference from which to assign our degrees of believe to the rules considered.

Finally, still the second requirement is to be addressed in order to be really assertive. It refers to the quantity or quality of knowledge expressed by the rule. As we exposed above, the more the cardinality, the more accurate the knowledge expressed. This is reflected in the sample as a lower prior probability of finding a higher order frequent item set, and a lower prior probability to find a rule involving higher cardinality features at a given confidence level. This fact is not yet explicit in this measures and therefore they present a significant bias in relation to cardinality.

The philosophy behind this approach is that, taking the certainty and uncertainty factors as a base expressing the quantity/quality of knowledge, a transformation can be applied in order to address this bias. We present a general expression for this transformation, wherein alternative and significantly different measures to *coverage*, *support* and *confidence*, can be derived.

The intention behind the new measures is to be as objective as possible.

As a result, the fact of *believing* what we *see* does not rely any more to an on/off switch, less or more accurately designed, as it happens with statistical measures. It does relay simply to the fact of observing that a certain rule appears among the first or the last positions in the ranked classification, being certainly sure that the former ones are expressing more useful knowledge than the latter, regardless of the actual reliability of the sample and regardless of the cardinality of the features involved. In this way we pretend to address the objections formerly exposed.

Chapter 4

Structural evidence functions

The basis of our measures suggests that meaningful transformations of this distances could be derived from linear or exponential approaches.

We forward that we have experimented with both and that they offer similar results. But from the very first beginning we chose the exponential one. The original reason for this was the idea that in a sample with white noise and given a particular antecedent, the value of the consequent should be normally distributed around a particular consequent, whenever they were engaged in a relation of dependence, and so had to be the deviation from the *ncd*. Later, this choice has shown to hold analytical properties that makes it still more preferable.

Hence, from an exponential perspective, a general form for this transformation would be,

$$Z(x) = k\, exp\left(\alpha\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{4.1}$$

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, μ is the correspondent mean ($1/s$ or $1/r$), and α and σ are two free parameters that will determine the final shape and scaling of this function. k is a normalizing factor, which is not meaningful as far as for this chapter, so we are going to drop it by now.

Recalling what we said in the previous section, we expect to tackle two different issues related with these parameters: (i) we expect to find an optimal balance in the way of measuring representativity and reliability, (refer to fig. 3.1), and (ii) we expect some kind of scaling in these measures related to the quantity/quality of knowledge expressed as a function of cardinality.

In fact, the combination $\sigma/\sqrt{\alpha}$ is what will determine both issues, and following our intuition, this should be a function of the certainty and uncertainty factors. But which relation is going to better handle both issues?

Furthermore, the certainty and uncertainty factors impose an asymmetry and this rises still another question, should or should not we regard this asymmetry in measuring the deviations from the equilibrium?

4.1 The QNormal distance distribution

It seems a matter of logic that shaping should hold asymmetry, while scaling should not.

Hence, we take different shaping factors at each side: the uncertainty factor $\sigma_- = U$ at the left, and the certainty factor $\sigma_+ = C$ at the right, and for simplicity we indicate $\sigma \equiv \{U, C\}$.

Regarding to the scaling factor, an intuitive way of handling asymmetry, is by directly combining the certainty and uncertainty factors in what we call the *knowledge factor*, given by,

$$Q = U\,C \tag{4.2}$$

Then taking $\alpha = ln\,(Q)$, renders the following expression for equation 4.1,

$$Z\,(x) = exp\left(ln\,(Q)\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{4.3}$$

where $\mu = U$ and $\sigma = \{U, C\}$, respectively at each side.

Further empirical validations of this expression show that still better balance is achieved by including an independent extra shaping parameter of $1/\,(2s)$. (We give further explanations of this question in section 5.3). Then, given that,

$$exp\,(ln\,(Q)\;A)\;=\;Q^A \equiv bxp\,(A)$$

expression 4.3 can be rewritten as,

$$Z\,(x) =\; bxp\left(\frac{1}{2s}\left(\frac{x-\mu}{\sigma}\right)^2\right) \tag{4.4}$$

where *bxp*, what we call the *knowledge factor exponential base*, is a self allowed notation derived from *exp*, (*natural exponential base*), with analogous meaning.

This is in fact a family of exponential functions with an obvious proximity to the *Normal distribution*, though we are not thinking of them as *probability density functions* but rather as distance measures, making sense only in the range $0 \leq x \leq 1$.

Therefore we refer to this family of functions as the *QNormal distance distribution*, $QN\,(U,\;\{U, C\})$, which are depicted in fig. 4.1.[1]

[1] Strictly speaking, this functions do not hold the formal properties of a metric distance functional (in particular, the triangular inequality does not make any sense here). They

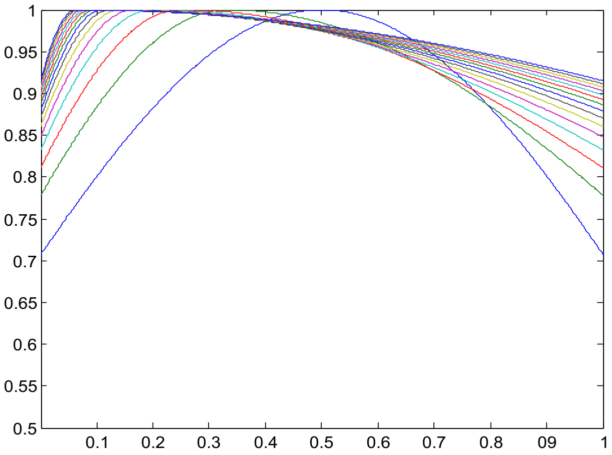


Figure 4.1: QNdd for $2 \leq s \leq 15$.

At the mean, the value of this function is 1, and at the boundaries, the function takes equal values, given by,

$$Z_z \equiv Z(0) = Z(1) = bxp\left(\frac{1}{2s}\right) \tag{4.5}$$

Following, we formalize the measures of representativity and reliability that can be derived from this general expression of distance, where we will use this value at the boundaries.

The final shape of this functions is determined by a special combination of U, C and Q: distances are measured relatively to the maximum possible deviation at each side, given respectively by U and C, and the scaling factor Q is a combination of U and C itself. The relation between them is depicted in fig. 4.2.

This particular combination renders an equilibrium, given at the mean, which is broken with a particular gradient that expresses two clashing facts: (i) the C factor expresses the idea that the more the cardinality, the more accurate the information given by the feature, therefore Q increases and the gradient decreases, so that more evidence must be seen in order to break the equilibrium, (ii) whereas the U 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 the gradient increases, making it easier to reach it.

should yet be regarded as deviations, but we feel more comfortable talking about distances because it fits better with the notions of *distance*/*proximity* to the *ncd*/*pmd*.

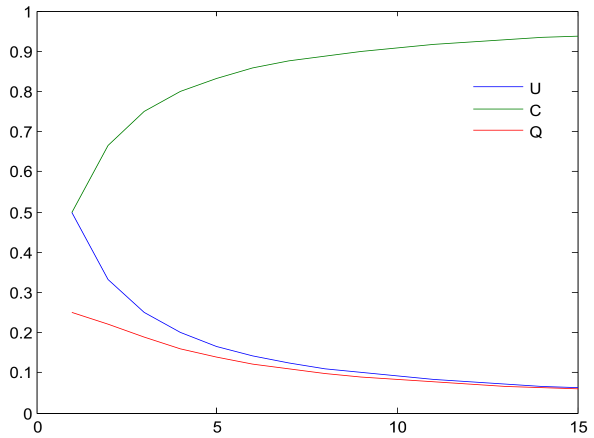


Figure 4.2: UCQ for $2 \leq s \leq 15$.

4.2 Presence

Applying the general expression given in 4.4 to the marginal distribution of feature X^p, we have,

$$\forall x_i^p \in X^p\,,\; z_i^p \equiv Z\left(\frac{n_i^p}{N}\right) = bxp\left(\frac{1}{2r}\left(\frac{\frac{n_i^p}{N}-\mu_r}{\sigma_r}\right)^2\right) \tag{4.6}$$

As exposed in chapter 2, our concept of coverage is regarded from the notion of contrastability rather then generality, and consequently representativity is an attribute (or sample) related concept, not an attribute-value one. This means that whenever we had a certain x_m^p with $n_m^p = N$, the representativity of that attribute should be zero, because no contrasting information about other values exist. Hence, our measure should be zero for any x_m^p with $n_m^p = N$, as well as for any x_z^p with $n_z^p = 0$.

Thus, we should fit the values of our measure in the interval $(0, 1)$. Also, it would be interesting to have normalized values. We can achieved this by combining 4.6 with the value at the boundaries given in 4.5, as it is shown in the following expression,

$$b_i^p = \frac{1}{r}\left(\frac{z_i^p - Z_z}{1 - Z_z}\right) \tag{4.7}$$

This function is an alternative and significantly different measure of *coverage*, which we call *presence*, and which is depicted in fig. 4.3 for different values of cardinality.

The total *presence* of feature X^p is then given by,

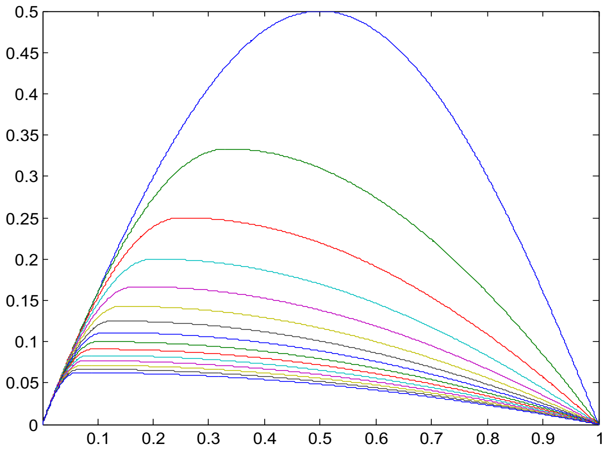


Figure 4.3: Presence function for $2 \leq s \leq 15$.

$$B^p = \sum_i b_i^p \tag{4.8}$$

with a maximum value of 1, given when all possible outcomes for the feature are equally covered.

As long as the marginal distribution of that feature moves away from the *pmd* in any direction, the value of *presence* decreases, vanishing at the boundaries.

4.3 Coherence

Applying the general expression given in 4.4 to the conditional distribution $(X^q \mid X^p)$, we have,[2]

$$\forall \left(x_i^p, x_j^q\right) \in (X^p, X^q)$$

$$z_{ij}^{pq} \equiv Z\left(\frac{n_{ij}^{pq}}{n_i^p}\right) = bxq\left(\frac{1}{2s}\left(\frac{\frac{n_{ij}^{pq}}{n_i^p} - \mu_s}{\sigma_s}\right)^2\right) \tag{4.9}$$

The exponential function inverts the initial measure, turning it into a proximity. But we are interested in measuring distances to the *ncd*, so we have to consider the opposite in this case.

As before, combining 4.9 with 4.5 in order to fit values into $(0, 1)$, and normalizing, we can derive an expression of reliability, given by,

[2] bxq stands for the knowledge factor exponential base for X^q

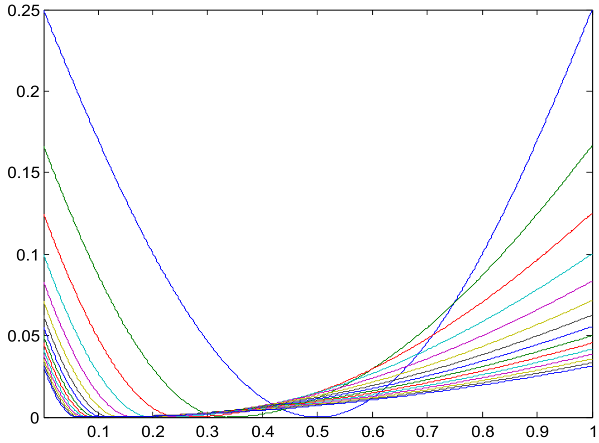


Figure 4.4: Coherence function with $r = 2$ and for $2 \leq s \leq 15$.

$$c_{ij}^{pq} = \frac{1}{r\,s}\left(1 - \frac{z_{ij}^{pq} - Z_z}{1 - Z_z}\right) \tag{4.10}$$

This function is an alternative and significantly different measure of *confidence*, which we call *coherence*. We depict it in fig. 4.4.

The total *coherence* of pattern $X^p \to X^q$ is then given by,

$$C^{pq} = \sum_{i,j} c_{ij}^{pq} \tag{4.11}$$

with a maximum value of 1, given when each subpattern is maximally coherent, as it is stated in the following definition,

Definition 3. 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* $1/\,(r\,s)$.

Obviously, it is an asymmetric measure, so that most of the time it will be $c_{ij}^{pq} \neq c_{ji}^{qp}$, and although this can not be directly grasped as the direction in which the cause-effect relation acts in the real domain, it may help at figuring out this matter for particular cases.

It is also notable, that antecedent and consequent can be sets of attributes. In this case, the measure refers to the relation of conditional dependence between features in the antecedent with respect to the consequent. That is, given $X^{pq} = \{X^p, X^q\}$ with $crd\,(X^{pq}) = (r\,s)$, and (X^c)

with $crd\,(X^c) = t$, the *coherence* of the conditional distribution $(\,X^c\,|X^{pq})$ is given by,[3]

$$z^{pqc}_{ijk} \equiv Z\left(\frac{n^{pqc}_{ijk}}{n^{pq}_{ij}}\right) = bxc\left(\frac{1}{2t}\left(\frac{\frac{n^{pqc}_{ijk}}{n^{pq}_{ij}} - \mu_t}{\sigma_t}\right)^2\right) \tag{4.12}$$

Therefore,

$$c^{pqc}_{ijk} = \frac{1}{r\,s\,t}\left(1 - \frac{z^{pqc}_{ijk} - Z_z}{1 - Z_z}\right) \tag{4.13}$$

4.4 Utility

Finally, combining the two former measures, we obtain the *utility* measure for the rule $x^p_i \rightarrow x^q_j$, which is given by,

$$u^{pq}_{ij} = c^{pq}_{ij}\,(b^p_i\,r)\,\left(b^q_j\,s\right) \tag{4.14}$$

The total *utility* of the pattern $X^p \rightarrow X^q$ is then given by,

$$U^{pq} = \sum_{i,j} u^{pq}_{ij} \tag{4.15}$$

with a maximum value of 1, given when *coherence* is maximal and *presence* for both features is perfectly equilibrated.

An example of the *utility* function for $x^p_i \rightarrow x^q_j$ with $(r = 2, s = 3)$ and being X^p in *pmd* is given in fig. 4.5.

4.5 Semantics of the utility function

This expression is not a strict measure of dependence. It intends to give an equable, impartial and equilibrated measure of *intensity and reliability of implication* in a relationship, taking into account its relative degree of representativity 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 *ncd*. 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. 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*.

[3] *bxc* stands for the knowledge factor exponential base for X^c

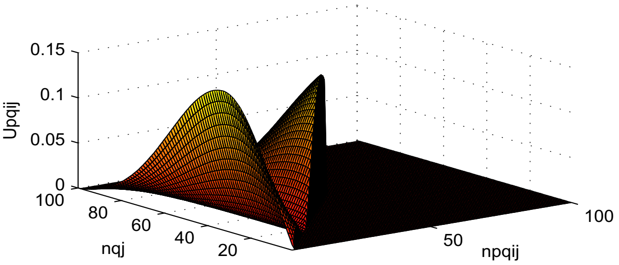


Figure 4.5: Utility function for $r = 2,\ s = 3,\ \text{and}, n_i^p = \frac{N}{r}$

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*).

By definition, *utility* is inversely related to the *total amount of uncertainty of the consequent given that the antecedent is known*, (see [5] for a related discussion). Even in the case of *independence*, we have,

$$U^{pq}\left(X^p \perp X^q\right) \geq 0$$

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 of statistical independence 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*.

4.6 Global and partial utilities

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.

Significant levels of utility are the following:

- *direct utility.*

  The set of *implications* of a feature is what we call its *direct utility.* It represents a set of conditional independence relations and is given by the sum of the *pattern utilities* with each one of its descendants (or consequents), that is,

  $$U_D^c = \frac{1}{crd\,(d)} \sum_{q \in d} U^{cq}\,, \text{ being } d \text{ the set of descendants} \tag{4.16}$$

- *inverse utility.*

  The set of *explanations* of a feature is what we call its *inverse utility.* It represents a set of conditional dependence relations and is given by the joint direct utility of its ascendants (or antecedents), that is,

  $$U_I^c = U^{Ac}\,, \text{ being } A \text{ the set of ascendants} \tag{4.17}$$

  Both partial utilities have significant implications regarding to *classification* and to *graphical modelling.*

- *total utility*

  Adding the two formers we get what we call the *total utility* of a feature, from which some conclusions can be derived related to *feature subset selection* issues.

- *undirected utility*

  Adding the *pattern utility* in both directions, we get the utility of the undirected relation among two features. In the next chapter we will show the great importance of this function with an example. It is given by,

  $$U^{pq/qp} = \frac{1}{2}\left(U^{pq} + U^{qp}\right) \tag{4.18}$$

- *global utility*

  And yet, adding the *undirected utility* of all dependencies considered in a particular model, we can get the utility at sample level, as a numerical reference of the quality of a sample with respect to that particular dependencies model.

  Recalling our first and second hypothesis, this would be a sort of measure of the *structural evidence* conveyed by the sample, given a particular dependencies model and given a particular partition, whenever discretization and clustering are involved. This allows us to choose among different alternatives without the need of checking them against any training or validation set.

An immediate striking practical application, is to implement this measure as a heuristic in a search algorithm in order to optimize a classification pattern for any feature. We give some examples of this application in chapter 6. It is worth mentioning that the *undirected utility* of that pattern should not be confused with its expected accuracy. Again, let's refer to the difference stated in the second hypothesis between structural an parametrical evidence. So far, we only have been talking about the former.

4.7 Utility function vs. Fischer Exact test

This section is an example of some ongoing work about comparing the *utility* function with other related statistical and descriptive measures of interestingness, as it is exposed in chapter 8. Though not thoroughly finished we consider it illustrative enough for including it here.

In fig. 4.6 we show a comparison of the *Fischer Exact Test* versus the *Utility function*, enhancing some important aspects about their behaviour with respect to different sampling situations.

We have chosen the Fischer Exact Test because, talking about dependence between features, it is commonly accepted to be one of the most reliable measures of interestingness, even with small and sparse samples.

Let's refer to a 2*2 contingency table, for a sample of size N, and row margin totals given by a and b, and column margin totals given by c and d. I'm showing only the sufficient values N, a and c.

In each graph we are plotting the values of the *Utility function* (the blue line) and the p-value given by a two sided *Fischer Exact test of homogeneity* (the green line), for all possible configurations of the contingency table, which are given in the x-axis as the value of the top-left cell (let's say n_{ac}). In order to make sense of the accept and reject regions of the null hypothesis we have also plotted a red line showing a significance level of 0.05.

The graphs in the first column show the effect of sample size in a situation of maximum information sampling scheme, that is, with equality in row and column margins.

In such situations, the Fischer Exact test gives the highest p-value (with a value of one) at the middle point, while the utility function gives its minimum (with a value of zero) at the same middle point. As far as here, that's obvious, as long as independence and minimum information are exactly the same joint distribution.

The first question to be noted is that while the Fischer Exact test acts as an on/off switch at some, more or less significant point, the utility function gives a continuous evaluation, assigning a value of 1 to both joint distributions expressing maximum certainty, and a value of zero to the joint distribution with minimum information.

We can also see that equal values of utility, falling in the accepting region

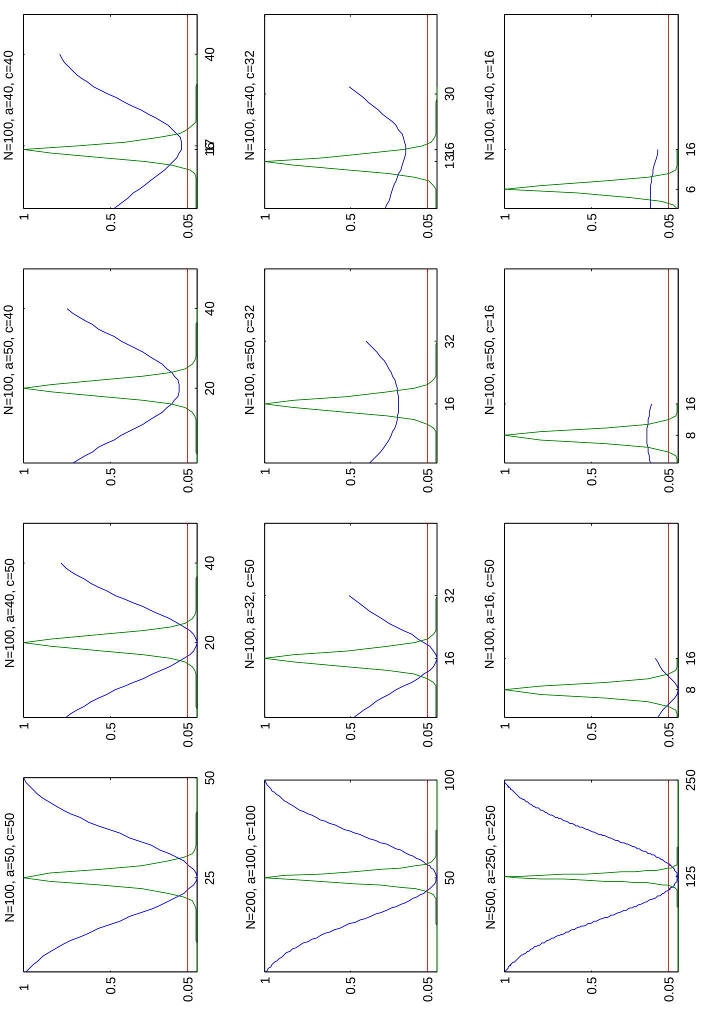


Figure 4.6: Utility versus Fischer Exact Test

of the null hypothesis for smaller sample sizes, are pushed to the rejecting region as the sample size increases, that is, as more evidence is available.

Right so far. But a major difference is that computational cost of the Fischer Exact Test may become unfeasible from a certain sample size on, specially when the total margins are so equilibrated and the set of possible contingency tables becomes huge. On the contrary, the computation of the utility function is a simple operation, quite independent of sample size, once the marginal and joint frequencies are known. This also holds for contingency tables of higher cardinalities or higher dimensions.

In the second column we show the effect of bias in the marginal distribution of the antecedent.

Remaining the consequent in marginal equilibrium, the joint distributions of independence and minimum information still coincide. But we can observe that while the Fischer test gives no extra information about that new situation, the utility function decreases the maximums, as long as knowledge conveyed by such joint distributions will never be as reliable as that coming from equilibrated marginals.

In the third column we show the effect of bias in the marginal distribution of the consequent.

In such situation, a joint distribution of zero information does not exist. There will always be some unbalance and consequently some conveying of knowledge. While the marginal bias is relatively low, independence and minimum information still coincide, though the minimum information is not zero anymore. The maximums are also lower than one. This clearly illustrates how the bias in marginal distributions, conditions by itself a minimum of conveyed knowledge, which is wiped away from the maximum, and also shown as a minimum greater than zero.

As the bias becomes greater, the independence and the minimum information joint distributions move away one from the other, and we reach a point, where the joint distribution becomes so tight, that there exists no configuration conveying more knowledge than that given in independence. That's what we can see in the lower graph where independence corresponds to maximum utility. This clearly illustrates our former assertion that the utility function is not a strict measure of dependence.

Finally, in the fourth column we show the effect of bias in both marginals.

In such cases, independence and minimum information hardly coincide. In the direction of the minimum of the marginals, some configurations can be found, for which the amount of information is less than that corresponding to independence. And it can even be the case, that for considerably high biased marginals, the minimum information point is falling in the rejecting region of the null hypothesis. That is, the Fischer test would accept the null hypothesis given a certain level of information, but would amazingly reject it for a lower level.

Chapter 5

Parametrical Evidence functions

Finally, the same approach holds yet another possible derivation from the parametrical point of view, which clearly explains what it is conceptually being done. This can be viewed as an alternative to the MLE or other parametrical learning methods.

5.1 Parametrical models

From the same QNdd function, we can derive parametrical models for the marginal and conditional distributions. As well as we dropped the normalizing factor while deriving the structural functions, in this case, we do need to normalize the QNdd function so as to have a probability density function.

A circular integration of 4.4 renders the following normalization factor,

$$k = 2\,\sigma\,\sqrt{\frac{-ln\left(Q\right)}{2\pi s}}$$

with different σ at each side of the mean. But, contrary to the *Normal* distribution, where one half of probability lies at each side of the mean, in our case, we want $1/s$ of probability at the left, and $\left(s-1\right)/s$ at the right, given precisely by σ. This renders equal normalization factors at each side.

Consequently, the normalized QNdd function, turns to a probability density function, given by,

$$\theta\left(x\right) = \sqrt{\frac{-2\,ln\left(Q\right)}{\pi s}}\; bxp\left(\frac{1}{2s}\left(\frac{x-\mu_s}{\sigma_s}\right)^2\right) \tag{5.1}$$

The depiction of this function is given in fig. 5.1.

By numerical integration of 5.1, we get the correspondent probability distribution function, depicted in fig. 5.2

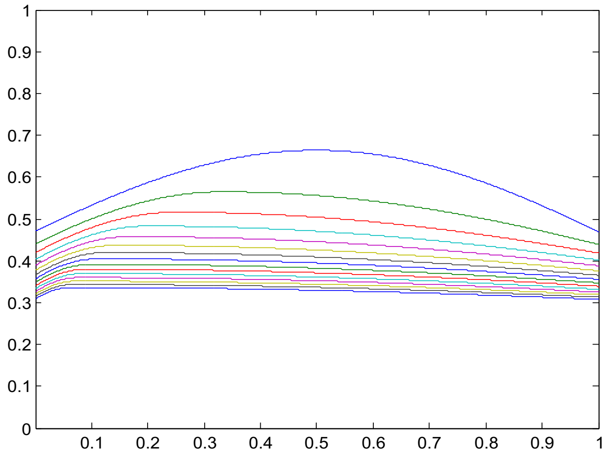


Figure 5.1: Probability density function for $2 \leq s \leq 15$.

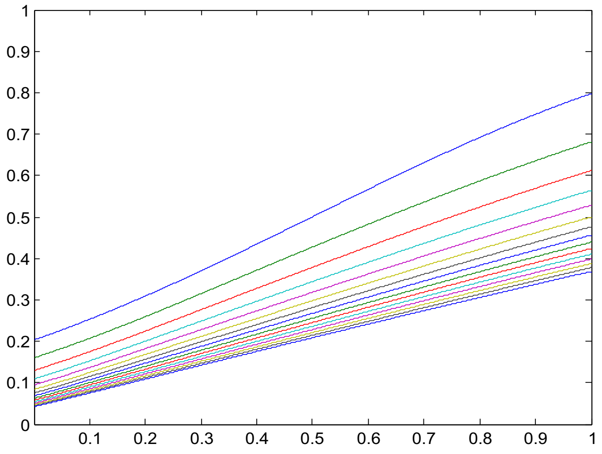


Figure 5.2: Probability distribution function for $2 \leq s \leq 15$.

So, instead of a single raw interpretation of support, we get a whole family of probability distributions as a function of the cardinality. This is of great importance, because it means that 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.

In the following we analyze some important properties of this family of distributions. An immediate conclusion that we can forward, is that this probability distributions express a more conservative understanding of evidence. Anyway, we have empirically verified, that this conservative trait turns to be usually enough in classification issues, 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 and methods already existent.

5.2 Complete family of parameters

The conservative trait of our parametrical models comes from the fact of considering the random nature of any sampling scheme. As an expression of the non 100% of reliability of the sample, this pdf gives non-zero values at the zero frequency and non-one values at the frequency one.

The non-zero values express the uncertainty associated to the fact of having no evidence of something. The non-one values express the uncertainty that should be regarded, in spite of having full evidence of something. So, the origin of this conservative trait is clear: some probability is deserved for the unseen cases.

And, as a direct consequence of the asymmetric normalization, the following relation holds:

$$1 - \theta\left(1\right) = \int_{1}^{\infty} \theta\left(x\right)\, dx = \left(s-1\right) \int_{-\infty}^{0} \theta\left(x\right)\, dx = \left(s-1\right)\, \theta\left(0\right) \qquad (5.2)$$

It is easy to figure out the meaning of this relation: seeing only one value, entails not seeing the other (s-1) values. So the uncertainty associated to both facts is the same.

A pretty useful direct consequence of this fact is that these parametrical models directly provide a full family of parameters.

5.3 Properties

Rather then properties, what we expose in this section is a desiderata of properties that our family of parametrical models should hold:

- Consistency

    1. $\sum_i^r \theta_i^p = \sum_i^r p\left(x_i^p\right) = 1$
    2. $\sum_j^s \theta_{ij}^{pq} = \sum_j^s p\left(x_j^q \mid x_i^p\right) = 1$

- Scalability

    1. $\sum_j^s \theta_{(ij)}^{(pq)} = \sum_j^s p\left(x_i^p, x_j^q\right) = p\left(x_i^p\right) = \theta_i^p$
    2. $\theta_{ik}^{pc} = p\left(x_k^c \left| x_i^p \right.\right) = \sum_j p\left(x_k^c \left| x_i^p x_j^q \right.\right) = \sum_j \theta_{ijk}^{pqc}$

Though stated separately, it is enough with proving either property 1 or 2 in both cases.

We have not yet verified whether our parametrical models hold these properties, but we have got some empirical indications that they indeed may hold them.

At that moment, this is probably the most important point of all our exposition. It would represent a kind of closure of our approach and a great contribution in order to endow it with some theoretical support and robustness. So, our immediate lines of research will go in the direction of finding a theoretical, (or empirical at least), prove of these properties.

In fact, the presence of the extra shaping factor of $1/\left(2s\right)$ in the QNdd comes from the empirical verification of part of these properties.

Chapter 6

Relational domain related issues

The aim of this chapter is to illustrate how the QNdd framework applies to pattern discovery in *classification* and *graphical modelling* problems. We do not pretend here to make any performance empirical comparative with other customarily applied methods. The main interest is only to highlight some particular questions of interest. Therefore, we will base our exposition on a toy synthetic example, exclusively designed for this purpose.

6.1 A toy example

Our example refers to a domain with three attributes $X = \{A, B, C\}$ involved in a relation of conditional dependence. All features are continuous, with values in the range (0,1). C is the class. A and B are independently distributed attributes and none of them gives any information about the class by itself, but when considered together they fully explain C.

In particular, there exists three levels of A and three levels of B, determined respectively by the boundary points (u_1, u_2) and (v_1, v_2), and there exist five levels (classes) of C, determined by the boundary points (w_1, w_2, w_3, w_4), and the relation holds the following set of rules,

$$\text{if } \ 0 \leq A < u_1 \text{ and } \begin{cases} \ \ 0 \leq B < v_1, \text{ then } \ 0 \leq C < w_2 \\ v_1 \leq B < v_2, \text{ then } w_1 \leq C < w_2 \\ v_2 \leq B < 1 \ , \text{ then } w_2 \leq C < w_3 \end{cases}$$

$$\text{if } u_1 \leq A < u_2 \text{ and } \begin{cases} \ \ 0 \leq B < v_1, \text{ then } w_1 \leq C < w_2 \\ v_1 \leq B < v_2, \text{ then } w_2 \leq C < w_3 \\ v_2 \leq B < 1 \ , \text{ then } w_3 \leq C < w_4 \end{cases}$$

$$\text{if } u_2 \leq A < 1 \;\text{ and } \begin{cases} \;\;0 \leq B < v_1, \text{ then } w_2 \leq C < w_3 \\ v_1 \leq B < v_2, \text{ then } w_3 \leq C < w_4 \\ v_2 \leq B < 1 \;, \text{ then } w_4 \leq C < 1 \end{cases}$$

In the following examples, we use different samples, drawn from this general joint distribution, by specifying different values for the boundary points. We pretend to show that we can find out the underlying pattern, under different conditions of data balancing.

The basis of our method is a heuristic search over the space of all possible partitions of the sample, based on a greedy optimization of the *undirected utility.*

We start the search with any initial partition. This is not relevant, so we use an equal width partition with an appropriate number of intervals for each attribute. Setting the initial number of intervals should be done in relation to the sample size N, because of the following: (i) too many intervals, with too few examples in each, may lead to a situation of excessive initial overfitting trapped in a local optimal, and (ii), too few intervals may difficult to reach the optimal partition.

Let's denote as r_a, r_b, and s, the respective cardinalities of A, B and C at any time of the search. Let's denote also $\pi = \{A, B\}$ as the set of ascendants of C, and $crd\,(\pi) = r_a\, r_b = r$.

Then the *undirected relation utility* is given by,

$$U^{\pi,c} + U^{c,\pi} = \sum_{l,k}^{r,s} u_{lk}^{\pi c} + \sum_{k,l}^{s,r} u_{kl}^{c\pi} =$$

$$= \sum_{l,k}^{r,s} c_{lk}^{\pi c}\, b_l^{\pi}\ (b_k^c\, s)\ + \sum_{k,l}^{s,r} c_{kl}^{c\pi}\ (b_k^c\, s)\ b_l^{\pi}$$

where,

$$b_l^{\pi} = \prod_{\rho \in \pi}^{m} \left(b_l^{\rho}\, r_{\rho}\right)^{\frac{1}{m}}$$

By definition of the *utility* function, that is the same as optimizing the conditional probability functions $P\,(C \mid A, B)$ and $P\,(A, B \mid C)$ together, while deserving a reasonable balance for the marginal distributions of each one.

The optimality of the final partition found should be guaranteed by a natural bias of the measure of *utility* toward the *minimum description length* principle. Anyway, this is subject to the efficiency of the searching algorithm itself. In this case we use a home-brewed simulated annealing, which we deem efficient enough to show the goodness and badness of this methodology, what is our main purpose.

The only parameters that must be fixed by the user are the number of intervals of the initial discretization of each feature. At the same time, this number sets an upper bound to it, and is expected to lower down as the optimization proceeds. If it does not, we should consider to repeat the optimization with larger values.

An important drawback is computational cost, being roughly of order quadratic with respect to the total cardinality involved in the search, given by (rs).

6.2 Discretization

Many of the existing discretization methods, ([7], [10], [11], [12], [14], [17], [19], [24], [25],[38]), are based on either one or both of the following: (i) independent discretization of each attribute, and (ii) supervised discretization.

Obviously, none of this assumptions applies to our case: the class has to be discretized itself and all attributes should be simultaneously discretized in order to find an optimal partition of the sample. That is what we refer to as simultaneous *domain sensitive*[1] discretization, formerly introduced in chapter 2.

Following, we give some examples to illustrate how our approach can deal with such a problem. We present three different cases corresponding to different situations of unbalance in data, as given in the following table,

	u_1	u_2	v_1	v_2
a) balance	0.33	0.66	0.33	0.66
b) slight unbalance	0.21	0.57	0.23	0.48
c) hard unbalance	0.12	0.31	0.16	0.35

In all cases, the class boundary values are equidistantly set. Unbalance in the marginal distribution of C is obviously determined by the joint distribution of A and B.

In the following figures we show graphical representations in which the black lines show, at the left, the underlying pattern used in the joint distribution, and at the right, the final pattern discovered. We use a color code to indicate the values of the class. The initial discretization is set to 10 equidistant intervals, equally for the three features.

[1] The concept of *domain sensitive* is inspired on the work by [25], which is close to ours, in the sense of stepping aside from the classical classification between super-

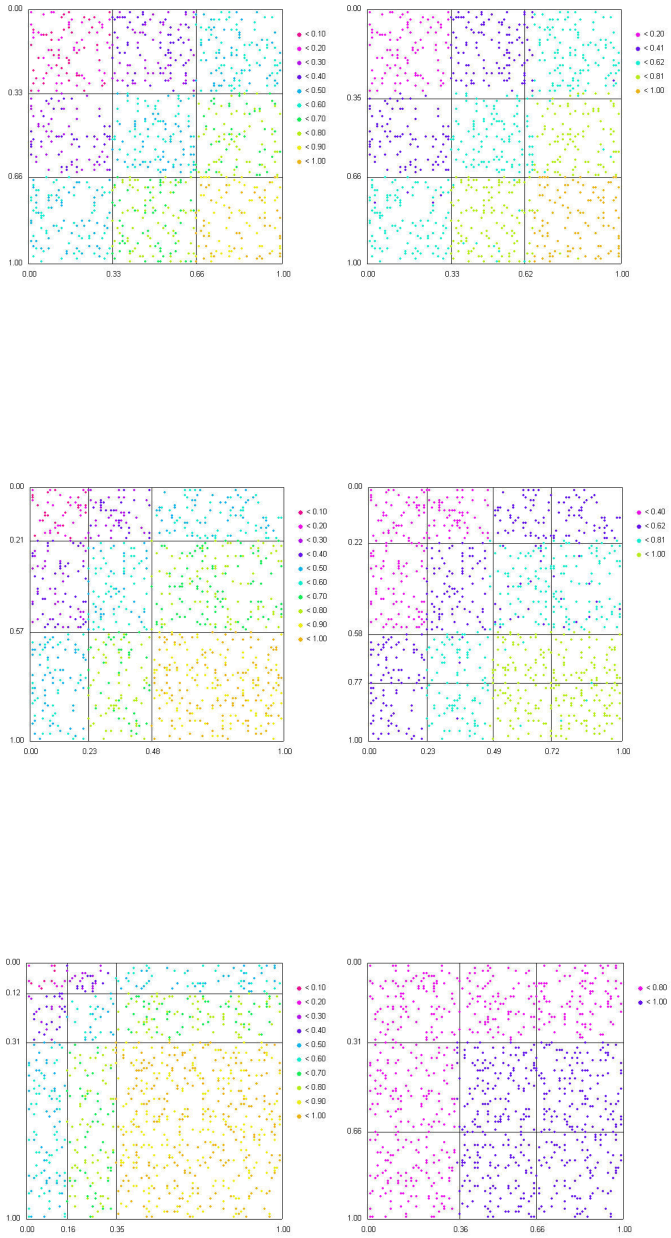


Figure 6.1: Discretization at different levels of unbalance

As it can be observed, the results are clearly affected by unbalance in data.

- balanced data (upper figure)

  The final pattern is quiet right. Only very small differences can be observed in the fine tunning of the boundaries. This is most probably due to some limitations of the searching algorithm. Further optimization would probably achieve a perfect model.

- slight unbalanced data (central figure)

  Some differences in the fine tunning of the boundaries are also observed here. Additionally, features A and B appear somewhat over discretized. In particular, for them both, the largest interval is split into two almost equal intervals. The reason is that, without any loss in *coherence*, a better balance in *presence* is achieved by this splitting.

  And finally, class c_1 has been merged with class c_2. The reason for this is that the evidence about this class is not enough to support it, with respect to the other classes. Holding this class, would contribute an amount of *coherence*, that would not compensate the loss in *presence* due to the unbalance generated in the marginal distribution of C.

- hard unbalanced data (lower figure)

  This case is more of the same of the previous, but obviously more extreme. Classes c_1, c_2, c_3 and c_4 are merged to one single class, and only c_5 is retained as originally. Again, the largest intervals of A and B are split in order to achieved better balance in their marginal distribution.

  Let's denote, that C is not split, though having as well some unbalance in its distribution. That is because in this case, such splitting would definitely damage coherence, and this loss would not be sufficiently compensated.

Far from being questionable, the differences in the final patterns under different unbalance, show exactly the same, not a different, behaviour. We have stated a certain will for believing what is seen in the samples, and the different patterns are just the consequence of it.

As a general observation, in all cases, the initial cardinality of features, set at 10, has been considerably reduced. This illustrates our assertion about the natural bias of our measure toward the MDL principle.

Whenever this strict conservative trend is excessive, and we have undesired results like merging of different classes, it is still possible to split the

vised/unsupervised optimization algorithms, though conceptually being a quite different approach.

sample as indicated by the pattern. In this way, each one of the subsamples gets more balanced. Then they can be further optimized as in a kind of zoom effect.

The former paragraph is not at all a kind of justifying argument. We believe that a deep truth is lying under it. We can not expect any accurate knowledge discovering tool to correctly find out all these patterns, in the same way as we can not expect to focus different size elements through a microscope.

Like discretization, clustering can be handled in the same way. The only difference is that in this case the search is over the space of all possible clusterings. We have had no time to prepare some examples, but there will be a specific section in this chapter illustrating this.

Other issues that should be extensively discussed in this chapter are the effects of sample size, noise, missing values and presence of outliers.

6.3 Feature Subset Selection

A direct consequence of the optimization process is that, whenever a feature is not relevant, its discretization tends to render a partition with one single interval, expressing in this way the uselessness of that attribute. Therefore, as a side effect of discretization, we get a sort of *feature subset selection.* This is yet another empirical confirmation of the natural bias that the measure of utility presents toward the MDL principle. (Similar references to this fact can be found in [12] and [24]).)

In order to illustrate this we present the examples shown in figures 6.2 and 6.3. The sample size is set to $N = 400$. All initial cardinalities are set to 7. And the boundary values are given in the following table,

	u_1	u_2	v_1	v_2
fig.6.2	0.24	0.56	0.10	1.00
fig.6.3	0.24	0.56	0.05	1.00

As it can be observed, A is slightly unbalanced and B is extremely unbalanced. In this situation, class c_5 does not exist, and B has quite of an irrelevant feature.

Even being so insignificant, the relevance of B is still detected in the first case. But a partition with only two intervals, would be excessively unbalanced. Consequently, B is over discretized, so as to render b_1 more significant with respect to the others.

In the second case B is definitely disregarded. This fact introduces some incoherence in the model. But the loss in *coherence* is less then the loss in *presence* of B due to considering such a small interval.

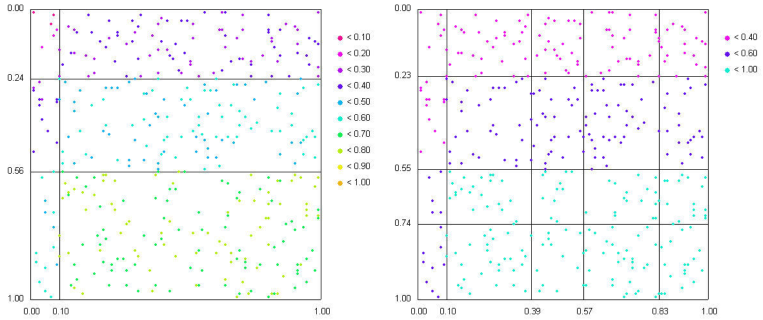


Figure 6.2: Relevance of B is still detected

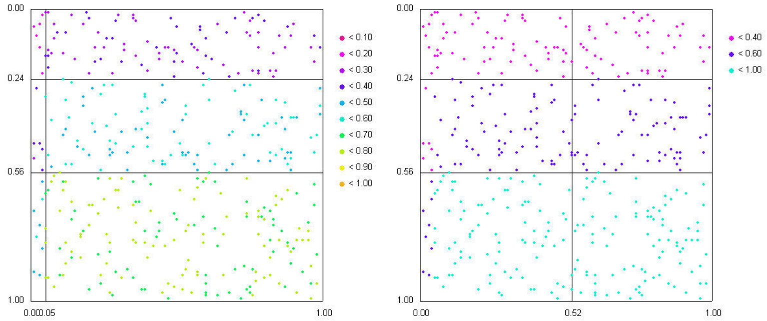


Figure 6.3: Relevance of B is disregarded

This may seem somewhat counterintuitive. Why are we disregarding such information? This is easy to understand if we forget that we know the real pattern. Then the reason is clear: though we are seeing this information, its evidential support is considered not enough so as to believe that, in later coming examples, it will remain exactly as the sample is showing. There is too much uncertainty associated to this evidence with respect to our will for believing in it.

Also, as it happened in the previous examples, evidential support for class c_1 is to less with respect to the other classes, hence it is merged with class c_2.

Chapter 7

Transactional domain related issues

In few words, the main conclusion drawn from the examples given in the previous chapter, is that it looks as if our approach assertively captures knowledge. As we have seen, this may have plenty of implications with respect to many relational domain related tasks. Then, it may as well be useful for transactional domain related tasks. So, a new line of future work refers to the study of how this framework fits in association rule mining issues, and what new contributions it may bring.

At first glance, *utility* may be regarded as an alternative interestingness measure that would allow to get a ranked classification of all rules extracted from any already existent rule mining algorithm. By definition, in doing that, it may contribute with issues like dealing with negative associations [33], redundant rules [2], and problems like over searching [30] and multiple comparisons [16] or multiple hypothesis testing [15], [26], [31], [36].

Still more interesting would be to develop pruning strategies based on our approach by itself. This is subject to the study of the algorithmic properties of *presence*, *coherence* and *utility*.

This is a line of research that has not yet been explored at all, and naturally, it is going to deserve a special treatment as part of our future work. In this sense we are going to follow the guidelines stated in [6], [20], [21], [22], [23].

7.1 Support and confidence thresholds

Any rule mining algorithm based on frequent set mining is subject to fixing the right support and confidence thresholds. This problem is a well known drawback of this framework and has raised a lot of controversy.

A direct approach to it comes from the family of probability distributions given in chapter 5. Given a desired threshold, our family of probability

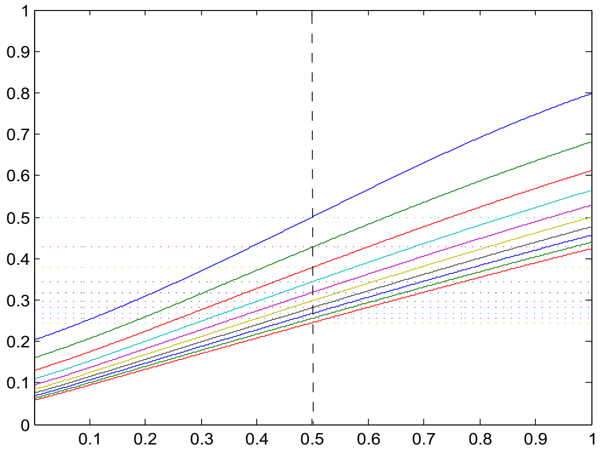


Figure 7.1: Support and Confidence thresholds for different k-itemsets

distributions provides different support and confidence thresholds for different levels of k-order itemsets. This contributes a very important part of the solution to this problem. We show this in fig. 7.1.

As it can be observed, given the desired threshold (50% for the example in fig. 7.1), lower values of support and/or confidence are fixed as the cardinality increases.

The conceptual reasons underlying this differences in the thresholds had already been introduced in section 3.4. They come from the following facts: (i) regarding to support, we are taking into account the lower prior probability of finding a k-order itemset in the sample, as a consequence of a higher cardinality, and, (ii) regarding to confidence, we are taking into account the lower prior probability of finding a rule with a given confidence level, as a consequence of a higher cardinality involved in it.

The former relates to a direct application of our parametrical models to other existing methods. But being *Presence* and *Coherence* alternative definitions to *Support* and *Confidence*, we can think about a direct approach using our measures.

In this case we have an extra advantage coming from the fact that, from our approach, each rule is considered beneath the framework of its whole pattern. Thus, not only we would have different thresholds for different k-itemsets, but the concept of threshold itself, turns into an interval of acceptable values. This is depicted in fig. 7.2.

For instance if we are willing to set a threshold of 50% of *presence*, this directly determines different upper and lower bounds of support for each k level. The upper bound avoids considering rules which have no significant counterpart in the pattern, complementing the knowledge they express as a counterfact.

The same holds for confidence. Fixing, again, a threshold of 50% of

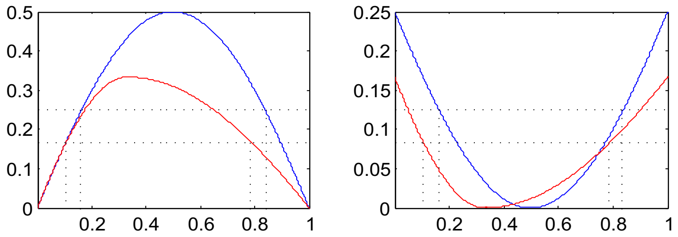


Figure 7.2: Support and Confidence upper and lower bounds for different k-itemsets, $k = \{2, 3\}$

coherence, different upper and lower bounds are determined for frequent k-itemsets of different levels, though in this case the valid intervals are the outside ones. The lower interval allows discovering rules expressing negative associations, which may as well be of interest.

All in all, it would have a helpful pruning effect at low levels, while allowing to still find valuable knowledge at higher levels.

Chapter 8

Planning

A whole new approach, with plenty of implications on many knowledge discovery related tasks, has been introduced. We are conscious of a great daring in doing it. Moreover, when ours is a quite conceptual posing, and while being based purely on intuition, a lack of supporting theory seems to be leaking wherever. We are aware of this, and we are aware of the great amount of work that must be done before a single one of these ideas can be respectably presented and, hopefully some day, commonly accepted. This is our objective.

Up till now, our aim has been mainly to set up a new point of view, from which some present directions appear to be questioned, and some overlooked topics are faced up with current methodologies. This has been by itself enough argument for our work.

While none of the proposals suggested here should be considered a closed question, our claim is to open a new door to further research, and we would like this to be regarded as our main contribution. Down from here, we deem some of our ideas promising enough to keep in this direction, hence deserving our proposal for this PhD Thesis.

8.1 Some conclusions

When not enough supporting theory is available, our last chance is to refer to empirical validation. Therefore, what follows is mainly based on the examples presented.

Though being a synthetic toy, they would be really challenging for many up-to-date methods. But regarding to our approach, and always from our opinion, they are illustrative enough so as to lead us to a single, necessary and sufficient, conclusion: our basis hypothesis hold, and the particular expression of the bias/variance dilemma within this approach, certainly states a particular will for seeing and believing.

Did we already manage to find out the right one? Probably not. Does exist a unique optimal will for seeing and believing? Probably neither. But we have shown that the quality of a sample can sure enough be measured, and that our approach can lead us to adeptly define this measure, taking into account such an elusive concept as it is the quality of information. Though not yet proved, the apparent properties shown by the parametrical models contribute additional traces that we may be in a right direction.

Yet more encouraging is the fact of realizing that our measure is really domain sensitive while capturing knowledge. If the measure works for discretizing, it may as well be useful for other knowledge discovery related tasks.

A different matter is the strict trait of our measure. We have seen how a class can be disregarded when unbalance in data exceeds its will for believing, and in some cases this may not be desirable. But from our opinion, before any subjective considerations it would be good to know where objectivity lies. This one should be our first goal.

Finally, let's say that our toy examples are in no way far from many of the real domain problems presented to the machine learning community, with continuous values and strong dependencies between features.

8.2 Future work

Our immediate next step must go in the direction of asserting the right measure of utility. We think to be very close to it. But at former times we already had this feeling, and we know that something unexpected is always there, ready to leap over.

A recent idea suggests that a good approach to it may come from the study of the analytical properties of our expressions. In this sense, the possible consistency and scalability of our parametrical models suggest very interesting ideas to work on. The idea of a closure as it is exposed in section 5.3 would be a great contribution.

Further validations about the certainty of our approach may come from the issues outlined in the following sections.

8.2.1 Analysis on the properties of our measures

Many efforts have been done in order to propose properties of interestingness measures, that may help to understand its characteristics and behaviour. Following the directions outlined in [6], [21], [22], [28], [34] a thorough study of the analytic and algorithmic properties of our measures should be done, in order to formally and conceptually place them among the most currently in use.

8.2.2 Analysis on the biases of our measures

We refer here to the problem of bias arising from different cardinalities. The certainty of our measure is tightly subject to the treatment given to the matter of the quantity and/or quality of information. Does our scaling adeptly reflect this?.

To the most of our knowledge, this issue has not been given so much attention. Some interesting studies exist about biases by cardinality in decision induction trees, [18], [37]. The main lines exposed in these works state a good reference in order to analyze the correctness of our treatment. The results of this analysis should be in accordance with the scalability property.

8.2.3 Comparative versus other measures of interestingness

A different and necessary analysis refers to the behaviour of our measure in different sampling situations, with respect to other measures of interestingness. In section 4.7 we have forwarded an example. But these different sampling situations does not refer only to sample size and unbalancing, but also to presence of noise, outliers and missing values. So, our example should be extended to those situations, as well as to other types of measures:

1. Statistical measures.
2. Entropy based measures
3. Support-Confidence based measures

8.2.4 Thorough comparative empirical validation

Finally, a thorough comparative empirical validation should be done based on some of the usual real-domain benchmarks.

8.3 Development schedule

By the year 2002, I had a good job as an informatics engineer, and I had time and money enough so as to decide joining the Artificial Intelligence PhD program at the LSI-UPC. I was already far from the thirties by that time. Since those days, I have been investing all my extra time and money on my education as a researcher, and on the ideas that have finally led to this work. Two years ago, I unfortunately missed my job. I had some indemnity from the job and a lovely daughter from my wife. Something told me then, that it was the time to take a sabbatical period for my daughter and my research. Now, money is gone and my daughter and my wife are fortunately still there. I have no contractual relation with the university, and no chance at all of applying for any scholarship because of the way these things work.

Therefore, as soon as I will have written the last word of this PhD Thesis project, I will have to run out urgently for a job. Hopefully, I will soon get one, and I will have one free hour a day, and energy enough, to keep on working in this PhD Thesis.

So, right now, I am in no way able to give any kind of development schedule. I could have skipped the former paragraph and give any devised schedule. But I rather prefer not to do it. Not because of any burden on my conscience. Rather because I want this section to be a claim against the fact that a forty years old student, stepping suddenly again into the academical environment, have no chance at all to get any kind of institutional help.

Bibliography

[1] AGRAWAL R., IMIELISKI T. AND SWAMI A. Mining Associations Between Sets of Items in Massive Databases. In *Proceedings of the ACM SIGMOD International Conference Management of Data*, pp.207-216, 1993.

[2] BALCAZAR J.L. Confidence Width: An Objective Measure for Association Rule Novelty. In *Proceedings of the Quality issues, measures of interestingness and evaluation of data mining models workshop* (QIMIE'09), pp.5-16, Bangkok, 2009.

[3] BASTIDE Y., PASQUIER N., TAOUIL R. STUMME G. AND LAKHAL L. Mining minimal nonredundant association rules using frequent closed itemsets. In *Proceedings of the 1st International Conference on Computational Logic*, London UK., pp. 972-986, 2000.

[4] BLANCHARD J., GUILLET F., BRIAND H. AND GRAS R. Mesurer la qualit des règles et de leurs contraposes avec le taux informationnel tic. *Revue des Nouvelles Technologies de l'Information*, E-2:287-297, 2004. Actes des journes Extraction et Gestion des Connaissances (EGC).

[5] BLANCHARD J., GUILLET F., BRIAND H. AND GRAS R. Assessing rule interestingness with a probabilistic measure of deviation from equilibrium. In *Proceedings of the 11th International Symposium on Applied Stochastic Models and Data Analysis (ASMDA'05)*, pp.191-200, ENST, 2005.

[6] LE BRAS Y., LENCA P. AND LALLICH S. On Optimal Rule Mining: A Framework and a Necessary and Suficcient Condition of Antimonotonicity. *In Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, (PAKDD'09)*, pp.705-712, Bangkok, April 2009.

[7] CHMIELEWSKI M.R. AND GRZYMALA-BUSSE J.W. Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning*, pp.319-331, 1996.

[8] Clark P. and Boswell R. Rule Induction with CN2: Some Recent Improvements in Machine Learning. In *Proceedings of the Fifth European Working Session Learning (EWSL'91)*, pp.151-163, 1991.

[9] Cohen W.W. Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning (ICML'95)*, pp.115-123, 1995.

[10] Dougherty J., Kohavi R. and Sahami M. Supervised and unsupervised discretization of continuous features. In *Proceedings of the twelfth International Conference on Machine Learning*, pp.194-202, 1995.

[11] Fayyad U.M. and Irani K.B. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pp.1022-1027, 1993

[12] Gama J., Torgo L. and Soares C. Dynamic discretization of continuous attributes. In *Proceedings of the Sixth Ibero-American Conference on AI* (1998), pp. 160169.

[13] Geng L. and Hamilton H.J. Interestingness measures for Data Mining: a Survey. *ACM Computing Surveys*, vol.38, no.3, article 9, September 2006.

[14] Ho K.M. and Scott P.D. Zeta: A global method for discretization of continuous variables. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pp. 191194, 1997.

[15] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, pp.65-70, 1979.

[16] Jensen D.D. and Cohen P.R. Multiple comparisons in induction algorithms. *Machine Learning*, 38(3), pp.309-338, 2000.

[17] Kerber R. Chimerge: Discretization for numeric attributes. In *National Conference on Artificial Intelligence*, AAAI Press, pp.123128, 1992.

[18] Kononenko I. On biases in estimating multi-valued attributes. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95)*, pp.1034-1040, Montreal, Canada, 1995.

[19] Kurgan L.A. and Cios K.J. CAIM Discretization Algorithm. *IEEE Transactions on Knowledge and Data Engineering*, vol.16, no.2, pp.145-153, February 2004.

[20] Lallich S., Vaillant B. and Lenca P. A Probabilistic Framework Towards the Parameterization of Association Rule Interestingness Measures. *Methodology and Computing in Applied Probability*, Springer Netherlands, vol.9, no.3, pp.447-463, September 2007.

[21] Lenca P., Vaillant B., Meyer P. and Lallich S. Association Rule Interestingness Measures: Experimental and Theoretical Studies. *Studies in Computational Intelligence (SCI)* 43, pp.51-76, 2007.

[22] Lenca P., Meyer P., Vaillant B. and Lallich S. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research.* vol.184., no.2, pp.610-626, January 2008.

[23] Li J. On Optimal Rule Discovery. IN *TKDE*, pp.460-471, Feb. 2006.

[24] Liu H. and Setiono R. Feature Selection via Discretization. *IEEE Transactions on Knowledge and Data Engineering*, vol.9, no.4, pp.642-645, July/Aug. 1997.

[25] Ludl M.C. and Widmer G. Relative Unsupervised Discretization for Association Rule Mining. In *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00)*, pp.148-158, September 2000.

[26] Megiddo N., and Srikant R. Discovering predictive association rules. In *Proceedings of the fourth international conference on knowledge discovery and data mining (KDD-98)*, pp.27-78. Menlo Park, AAAI Press, 1998.

[27] Padmanabhan B. and Tuzhilin A. Small is beautifull: Discovering the minimal set of unexpected patterns. In *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining* (KDD 2000), pp.54-63, Boston MA, 2000.

[28] Piatetsky-Shapiro G. and Matheus C. Discovery, analysis, and presentation of strong rules. *Knowledge Discovery in Databases*, pp.229-248., 1991.

[29] Quinlan J.R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[30] Quinlan J.R., and Cameron-Jones, R.M. Oversearching and layered search in empirical learning. In *IJCAI95*, pp.1019-1024., 1995.

[31] Shaffer J.P. Multiple hypothesis testing. *Annual Review of Psychology*, 46, pp.561-584, 1995

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

[33] Silverstein C., Brin S. and Motwani R. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. *Data mining and Knowledge Discovery*, 2(1), pp.39-68, 1998.

[34] Tan P., Kumar V. and Srivastava J. Selecting the right interestingness measure for association patterns. In *Proceedings of the 8th International Conference on Knowledge Discovery and data Mining (KDD'02)*, pp.32-41, Edmonton, Canada, 2002.

[35] Suzuki E. Interestingness measures - Limits, Desiderata, and Recent Results - In *Proceedings of the Quality issues, measures of interestingness and evaluation of data mining models workshop (QIMIE'09)*, pp.1-3, Bangkok, 2009.

[36] Webb, G.I. Discovering Significant Patterns. *Machine Learning*, 68, 1, pp.1-33, Jul. 2007.

[37] White A.P. and Liu W.Z. Bias in Information-Based Measures in Decision Tree Induction. *Machine Learning*, 15, pp.321-329, 1994.

[38] Yang Y. and Webb G.I. Weighted proportional k-interval discretization for naive-Bayes classifiers. In *Submitted to The 2002 IEEE International Conference on Data Mining*, 2002.