A GENERAL CRITERION FOR MEASURING QUALITY OF CONCEPT DESCRIPTIONS

Francesco Bergadano[1], Stan Matwin[2],
Ryszard S. Michalski, Jianping Zhang

Artificial Intelligence Center
George Mason University,
Fairfax, VA 22030

MLI 88-8
TR-13-88

October 1988

[1] On leave from University of Torino, Italy

[2] On leave from University of Ottawa, Canada

This paper is an improved and updated version of the paper presented at 3rd European Working Session on Learning, Aberdeen, Scotland, Oct. 3-5, 1988

A GENERAL CRITERION FOR MEASURING QUALITY OF CONCEPT DESCRIPTIONS

ABSTRACT

An important aspect of any learning method is an evaluation of the learned knowledge, in particular, an evaluation of the plausibility and usefulness of concept descriptions that are being created. This paper presents a new, general method for evaluating concept descriptions.

The method applies not only to the conventional logic-style concept descriptions, but also to *two-tiered* descriptions that characterize imprecise and/or context-dependent concepts, such concepts are called as flexible. In such descriptions, the first tier specifies typical and idealized concept properties, and the second tier describes the variability and allowed modifications of these properties in different contexts, and exceptional cases. Another novel feature of the measure is that it takes into consideration the *typicality* of cases covered by the description.

In the proposed measure, the quality of a concept description depends on three major criteria: the *accuracy*, the *comprehensibility* and the *cost*. To illustrate the measure, two alternative descriptions of the concept "chair" are evaluated . This work was done in the context of research efforts to develop a general method for learning two-tiered concept description.

ACKNOWLEDGEMENTS

The authors thank Dr. Gail Thornburg, Joan Elliott and Ken Kauffman for useful comments and criticism.

This research was done in the Artificial Intelligence Center of George Mason University. Research activities of the Center are supported in part by the Defence Advanced Research Projects Agency under grant, administered by the Office of Naval Research, No. N00014-87-K-0874, in part by the Office of Naval Research under grant No. N00014-88-K-0226, and in part by the Office of Naval Research under grant No. N00014-88-K-0397. The first author was supported in part by the Italian Ministry of Education (ASSI) project, and the second author was supported in part by the NSERC grant No. A2480.

1 INTRODUCTION

Inductive inference is one of the basic strategies of learning. Given examples, background knowledge, and optionally, an initial concept descriptions, this strategy hypothesizes a general concept description. Usually a large number of general descriptions can be generated for any incomplete set of examples and/or initial concept descriptions. To choose among candidate hypotheses one needs a criterion for preferring one description over the other.

This paper proposes a general measure for evaluating the quality of concept descriptions. The measure applies to descriptions produced by an automated learning system, as well as to descriptions created by a human. The measure was specifically designed for evaluating descriptions of flexible concepts. Such concepts are described using a two-tiered concept representation (Michalski, 1987). The results presented are part of a larger project on the development of systems for inductive learning of two-tiered concept representations.

Inductive inference is not truth-preserving, but falsity-preserving. Thus, the correctness of descriptions generated is uncertain. In evaluating these descriptions several factors can be taken into consideration. One is the relationship between the learned description and initial examples. Such a relationship may, for instance, show the completeness and consistency of descriptions with regard to the examples from which it was generated.

Another factor is the predictive power of the description, i.e., a measure of the description performance on new examples. One may also consider the *simplicity* of a description, and the ease of explaining it in terms of concepts already known to the learner. These two factors together affect what we call here the *comprehensibility* of a description. Finally, one may take into consideration the *cost of measuring* variables and terms in descriptions, as well as the *cost of storing,* and evaluating the description in order to predict new facts.

The problem of evaluating descriptions is not new, and a number of measures of description quality have been developed in the past. Some of them concentrate solely on the aspect of completeness and consistency (e.g. Mitchell, 1977). Other measures include also other criteria, such as the simplicity and the cost of evaluating the learned descriptions (Michalski, 1973).

A common assumption is that the simplicity of a hypothesis, and its performance on new facts are primary factors in evaluating it. To determine the performance of a description on new facts requires that new facts are available. Therefore, such a criterion is not applicable during hypothesis learning, when the learner needs to choose among competing candidate hypotheses before testing them on new data. This paper is concerned with such a situation, that is, with determining the quality of a description during the process of learning.

The simplicity of a hypothesis has been traditionally the major criterion for choosing among competing hypotheses (e.g., Kemeni, 1953). Popper (1968) pointed that simpler descriptions are easier to refute, and therefore are preferable. Pearl (1978) indicated that there is a connection between the simplicity and the probability of correctness of a hypothesis. Many evaluation criteria related to simplicity have been employed in automated learning systems (e.g., Michalski, 1980; Bergadano, Giordana, Saitta, 1988).

Broader aspects of the problem of what should be the preference criterion for judging competing inductive hypotheses are discussed in (Mitchell, 1980; Michalski, 1983; Utgoff, 1986; Michalski, Carbonell, Mitchell, 1983) Recently, Medin, Wattenmaker and Michalski (1986) presented results of psychological testing which indicated that humans use not only simplicity but also other criteria for selecting inductive hypotheses.

The problem of defining adequate preference criterion is a fundamental issue still unresolved in machine learning. This paper presents a quality measure of a description that combines the above mentioned three factors: *accuracy*, *comprehensibility*, and *cost.*

The accuracy of a concept description reflects the degree to which the description relates to the concept it describes. In the case of concept learning from examples, accuracy depends on the completeness and consistency of the description with regard to learning examples. It also depends on the typicality of the examples it covers, and the justification that can be constructed for the description. If the description can be plausibly justified in terms of the domain knowledge, the confidence in its correctness increases.

When a two-tiered concept description is used (Michalski, 1987), the quality of a description needs to relate to both the explicit representation (1st tier), and the implicit representation (2nd tier). Thus, quality has to take into consideration the exceptions from the base concept, and its admissible transformations.

The proposed quality measure has two novel aspects. First, it takes into account a number of different criteria, such as the degree of consistency and completeness of the description, the typicality of examples covered by it, its comprehensibility and the cost of storing and evaluating the descriptions. Second, the measure can be applied to concepts represented in a two-tiered form.

2 TWO-TIERED CONCEPT REPRESENTATION

Before we define the proposed description quality measure, let us first describe basic ideas underlying the two-tiered concept representation. In the traditional representation ("one-tiered"), any concept is defined by specifying basic features that cover all instances of the concept. It is often assumed that these instances can be described by a single conjunct.

In a more general approach a description consists of several conjuncts linked by disjunction. Each such conjunct contributes to the accuracy of the description, depending on how many examples it covers or explains. If these conjuncts are put in the order of decreasing coverage of examples (Michalski et al., 1986), the obtained description evokes an analogy with the Taylor series expansion of a function. In such an expansion, consecutive terms contribute a decreasing amount to the total value of the function. A concept description that includes rare or exceptional events will typically have a number of conjuncts that cover only small number of events. A complete description may therefore be overly complex, difficult to comprehend and have high cost (as defined below).

To deal with this and related problems, Michalski (1987) has proposed a two-tiered concept representation. The complete concept description is split into two-parts: the Basic Concept Representation *(BCR)* and the Inferential Concept Interpretation *(ICI)*. The BCR defines the concept simply and explicitly by characterizing the typical or ideal concept cases either in terms of attributes observed in the examples, or in terms that are constructively learned during concept formation. The prototypical instances of the concept can therefore be classified by simple matching with the BCR.

Anomalies, exceptions and context-dependent cases are handled by the ICI, which involves a reasoning process. The ICI deals with exceptions by inferring that they are instances of the concept (concept *extending*), or that they ought to be excluded from the description in the BCR (concept *shrinking*). The ICI uses production rules which may be deductively chained. A simple form of ICI is to define a certain similarity (or distance) measure to classify examples that are similar to those covered by the BCR. (Such an approach is used in *flexible matching* described in Michalski, Mozetic, Hong, Lavrac, 1986.)

Let us illustrate the idea of two-tiered representation with the concept of *chair*. The dictionary (Random House) gives the following definition:

> 1. a seat, esp. for one person, usually having four legs for support and a rest for the back and often having rests for the arms. 2. a seat of office or authority. 3. a position of authority, as of a judge, professor, etc. 4. the person occupying a seat of office, esp. the chairman of a meeting. 5. see electric chair. (...)

The description indicates several meanings, but does not tell when each meaning is applicable. It makes no distinction between the typical meaning and context-dependent meaning. It is rather hard to comprehend, and it is incomplete. A two-tiered representation of the chair concept could have the following form:

BCR:

A piece of furniture typically used for sitting by one person. Usually consists of a seat, four legs, and a backrest. (A picture of a typical chair, or a description of the relationship among the parts may be included).

ICI:

The number of legs may vary from 1 to 4

the shape, the size, the color and the material of all components can vary as long as the function defined in the BCR is preserved

chair without the backrest ---> stool rather than chair

chair with arm-rests ---> chair specializes to armchair

context = museum exhibit --> chair is not used for seating any more

context = toys --> Dimensions can be much smaller, but other physical properties are preserved. Does not serve for sitting by normal persons, but by correspondingly small dolls.

context = execution --> specializes to electric_chair

This simple example illustrates several important features of the two-tiered representation. Typical examples match the BCR, and therefore it is easy to identify them. The ICI involves metaknowledge, e.g. showing which properties in the BCR are crucial and which are not; and context-dependent knowledge, showing how the properties change in different contexts. In general, contexts can be hierarchically organized, and the ICI inference rules may chain, (although it is not shown in this simple example).

We argue that the "quality" of the two-tiered representation is higher than the quality of the dictionary definition, if used in an AI system. First, the accuracy is improved, since the two-tiered description is more complete and consistency has not changed. Second, comprehensibility is somewhat greater, because the prototypical properties of the chair are separated from its possible modifications and specializations.

A few systems that generate and use two-tiered representations have been described in the literature (Michalski et al., 1986, Bergadano et al., 1988, Bergadano, Giordana, [to appear]). Generally, two-tiered descriptions tend to be simpler, easier to understand and more efficient to use than conventional ones. They may also have higher performance on the testing set. In the systems developed so far, the ICI has been implemented in the form of flexible matching (Michalski et al., 1986). Such a matching performs only a similarity-based determination of the degree of match, but no rule-based reasoning. It is relatively fast, but not very deep. An improvement in the quality of a description is therefore measured only by the improvement in the first tier.

As with any measure of description quality, the final evaluation is only possible with the use of a testing set. However, because the testing cases are assumed to be unavailable during learning, it is necessary to evaluate them without testing cases. This paper defines a measure of the quality of concept descriptions without the benefit of testing cases. General requirements for such a measure are specified, and a specific measure is defined and illustrated by an example.

3 CRITERIA AFFECTING QUALITY OF CONCEPT DESCRIPTIONS

As mentioned earlier, the quality of a concept description is influenced by three basic characteristics: the accuracy, the comprehensibility, and the cost. This section discusses these three components, and describes a mechanism for combining them into a single measure.

3.1 Accuracy

Accuracy represents the description's ability to produce correct classifications. The basic criterion that relates to accuracy is the completeness and consistency of the description with respect to the learning events (Michalski, 1973, Mitchell, 1977, Michalski, 1980). In order to achieve completeness and consistency in presence of noise, one may have to generate overly complex and detailed descriptions. Such descriptions, however, may not perform well in future cases and examples. This is the well known phenomenon of overfitting (S. Watanabe, 1969; E. Sturt, 1981).

If a description is incomplete and inconsistent, the relative number of uncovered positive examples and the relative number of covered negative examples provide important information for evaluating its quality. If the description is also sufficiently general and does not depend on the particular characteristics of the learning events, these measures can be a meaningful estimate of the accuracy of the description.

Completeness and consistency of a two-tiered description brings up additional requirements: a good representation should cover the typical examples explicitly, and the non-typical ones implicitly. Moreover, the coverage of typical negative examples in the BCR is particularly detrimental to the quality of the representation. This is important to accuracy because the BCR is mainly obtained or justified by the learning events. Therefore, one can be confident in the information contained in the BCR only if a sufficient number of examples are available, or if the examples are typical or representative for the domain.

On the contrary, the ICI, being generated by experts or derived from the available domain knowledge, is appropriate to handle rare or exceptional events. In evaluating the accuracy of a two-tiered representation, we have to take into account the fact that degree of confidence in the results of inference decreases when going from deduction to induction (Michalski, 1987).

The above requirements are met by making the degree completeness and consistency dependent on the typicality of the covered examples and on the way these examples are covered. We assume that an expert can provide the typicality value of examples at the time they are presented to the system responsible for creating the initial description.

The degree of generality is also related to accuracy, since it affects predictive power. Given the same degree of completeness and consistency, a learning system should prefer maximally specific characteristic descriptions or maximally general discriminant descriptions. Characteristic descriptions are better if more specific, because they theoretically distinguish a given set of concepts from the set of all the other possible concepts. For example, if we want to characterize the concept of a cat, we will prefer the description "small feline" rather than the description "animal", since the first one is more specific. On the contrary, if we were to distinguish between cats and dogs, "feline" will be a better discriminant description of "cat" than "small feline", since it is more general and still sufficient. The number of different events that the description could possibly cover may be used to measure generality (Michalski, 1983).

The accuracy of a description can also be predicted by trying to justify the inductive hypotheses on the basis of general and domain knowledge. Such knowledge can be used to evaluate expressions. It may supply a measure of *importance* for the descriptors in the language, so that expressions containing better descriptors will be chosen.

3.2 Comprehensibility

Comprehensibility of the acquired knowledge is related to subjective and domain-dependent criteria. Because an AI system is often supposed to supply advice to humans, knowledge used by such a system should be understandable by human experts. A black box classifier will not be accepted by experts as a help in their work. Therefore, knowledge acquired by a learning system should be related to terms, relations and concepts used by experts, and should not be syntactically too complex. This is called the *comprehensibility* criterion (Michalski, 1983).

There is no, however, established measure of comprehensibility of a description. In our method, we will approximate it by representational *simplicity*, that evaluates the syntactic simplicity of the description's expression by counting the number of operators involved. The complexity of operators has been taken into consideration also. For more detail, see sec. 4.4.

3.3 Cost

The *cost* captures the properties of a description related to its storage and use (computational complexity). Other things being equal, descriptions which are easier to store and easier to use for recognizing new examples are preferred. When considering the cost of a description, two characteristics are of primary importance. The first one is the cost of measuring the values of variables occurring in the description. In some application domains, e.g., in medicine, this may be a very important consideration. The second one is the computational cost of evaluating the description. Again, certain applications in real-time environment, e.g., speech or image recognition, may impose constraints on the evaluation time of a description. The cost (approximated by computational simplicity) and the comprehensibility (approximated by representational simplicity) are usually related to each other, but in general there are different criteria. In sec 4.5, we will give more detailed description.

3.4 Combining Several Criteria

The above criteria need to be combined into a single evaluation procedure that can be used to compare different concept descriptions. A possible solution is to have an algebraic function that, given the numeric evaluations of single criteria, produces a number that represents their combined value. Examples of such functions are multiplication, weighted linear sum, maximum/minimum, or t-norm/t-conorm (Weber, 1983). Although these approaches are often appropriate, they have certain disadvantages. Firstly, they usually combine a set of heterogeneous evaluations into a single number, and the meaning of this final number is hard to interpret for a human expert. Secondly, they may force the system to evaluate all the criteria, even if it would be sufficient to compare two given descriptions on the basis of the most important one, if one is much better than the other. Thus, they may be overly complex if a large number of criteria have to be evaluated.

Finally, because the goal of the evaluation is to determine the "best" description, it is sufficient just to rank the candidate descriptions. Therefore, there is no necessity to assign some specific value of "quality" to them. An attractive method that solve avoid the problem mentioned above is to use the *lexicographic evaluation functional* (LEF) (Michalski, 1972, Michalski, 1983). The LEF gives a general measure of description quality by combining accuracy, comprehensibility and cost into a ranking function of description. The general description quality (GDQ) measure is thus defined as:

$$\mathrm{GDQ(description)} = <(\mathrm{Accuracy},\tau_1), (\mathrm{Comprehensibility},\tau_2), (\mathrm{Cost},\tau_3)>$$

where τ_1, τ_2, and τ_3 are tolerance thresholds (which will be discussed later). In this evaluation scheme, the criteria are ordered according to their importance, and a tolerance threshold ($\tau_i \in$ [0..100%]) is associated with each criterion. Given a set of descriptions, the LEF determines a set of most preferable descriptions in the following ways.

First, all descriptions are evaluated from the viewpoint of *accuracy*, and those which score best, or within the range defined by the threshold τ_1 from the best, are retained. Next the retained descriptions are evaluated from the viewpoint of *comprehensibility* and reduced similarly as above, using tolerance τ_2. Finally the same process repeats from viewpoint of *cost*. All descriptions retained now are equivalent from the viewpoint of the LEF. It worth to mention that whenever only one description is retained, the evaluation terminates.

The LEF evaluation scheme is not affected by the main problems which affect algebraic functions which we have discussed above. The importance of a criterion depends not only on the order in which it is evaluated in LEF evaluation scheme, but also on its tolerance. It may be difficult to determine this tolerance. If the tolerance is too small, we have very little chance of using the other criteria. If the tolerance is too large, some important criterion might be underestimated. Furthermore, in the case of a large tolerance, many descriptions might be equivalent under the LEF evaluation scheme. In order to avoid this problem, the LEF measure can be extended in the following way: LEF is first applied with larger tolerances, in such a way that all the relevant criteria are taken into account. If the comparison still results in a tie, a Weighed Evaluation Functional (WEF) is used to combine the measures (i.e. the description having the maximum weighted sum of the measures is preferred). The weights for WEF are determined by user.

The above criteria can also be applied to two-tiered descriptions. The accuracy of the acquired knowledge does not depend only on the explicit information, but also on implicit reasoning abilities. Inferential Concept Interpretation also affects cost, since it allows the performance system to use a simpler BCR, and reason about special details only in exceptional cases. Finally, the comprehensibility of a two-tiered representation must be carefully evaluated, since one of its implied goals is to state a clear and simple concept description in the BCR and to account for meaningful special cases through a reasoning process.

4 THE QUALITY MEASURE

In the previous section, we proposed a general framework for evaluating the quality of concept descriptions. In this section, we present a more precise and slightly simplified measure based on the scheme mentioned above:

$$\text{Quality(description)} = <(\text{Accuracy}, \tau_1), (\text{Comprehensibility}, \tau_2), (\text{Cost}, \tau_3)>$$

which is evaluated using LEF/WEF introduced in the previous section.

4.1 Types of Description Matching

An event can be covered by a two-tiered description through the following three types of matching:

1. **Strict matching**: the event matches the BCR exactly that is satisfies the conditions stated in BCR, in which case we say that the event is S-covered,
2. **Flexible matching**: the event matches the BCR through a flexible matching function, and we say the event is F-covered.
3. **Deductive matching**: the event matches the concept through deductive reasoning by using the ICI Rules, and we say the event is D-covered.

These three sets of events are mutually exclusive. S-covered events are *explicitly* covered, and F-covered and D-covered events are implicitly covered.

4.2 A Measure of Completeness Consistency in the Case of Examples of Different Typicality

Before we define *accuracy,* we first introduce *Typicality-dependent Completeness* (TYCOM) and *Typicality-dependent Consistency* (TYCON), and we discuss some issues related to these concepts.

The degree of completeness of a description is defined as the ratio of the number of positive examples covered by the description and the number of total positive examples supplied. The degree of consistency of a description is defined as 1 minus the ratio of the number of negative examples covered by the description and the number of total negative examples supplied. In the case that a description is incomplete and inconsistent, The degree of completeness and consistence is a important measure of the quality of a description. Such a measure has to take the typicality of examples covered and not covered by the description into consideration. These degrees are called TYCOM and TYCON respectively.

In general, descriptions that cover many typical positive events are most preferred. The degree of completeness should therefore be proportional to the typicality of the events covered. Moreover, if negative events are covered, the degree of consistency of the description should be inversely proportional to the typicality of the negative events covered.

As mentioned before, it is preferred that the typical events be covered by the BCR, and non-typical, or exceptional events be covered by the ICI. In fact, the BCR is inductively learned from the events provided by user, and it is more reliable when the learning events are typical. The rules for ICI, on the contrary, are inherited from higher level concepts, or provided by a human expert, and rely more on general and domain knowledge. Typically the ICI plays the most important role when dealing with special or rare cases. For these reasons, typical positive events that are explicitly-covered should contribute to completeness more than those typical events that are implicitly-covered. And vice-verse, nontypical positive events that are implicitly-covered should contribute to completeness more than those non-typical positive events that are explicitly-covered.

Furthermore, because ICI rules are obtained from background knowledge or from a human expert, they are more reliable than the flexible matching function. Consequently, a positive D-covered event should contribute to completeness more than F-covered. We may also observe that flexible matching is not very useful for exceptions whose typicality is very small. This is because that flexible matching is only useful for the events that are similar to the typical events. For example, a typical chair has four legs. Some chairs have three or five legs, flexible matching works fine with these chairs which have three or five legs. But it does not woke with a wheel chair. In order to recognize a wheel chair, some ICI rules are needed. A similar argument holds for consistency.

Now, we define the typicality-dependent completeness (TYCOM) of a description:

$$\mathrm{TYCOM} = \frac{\sum\limits_{e^+ \text{ is S-covered}} w_s * \mathrm{Typicality}(e^+) + \sum\limits_{e^+ \text{ is F-covered}} w_f * \mathrm{Typicality}(e^+) + \sum\limits_{e^+ \text{ is D-covered}} w_d * \mathrm{Typicality}(e^+)}{\sum\limits_{e \in \mathrm{PosCov}} \mathrm{Typicality}(e)}$$

Typicality-dependent consistency (TYCON) of a description is defined as follow:

$$\mathrm{TYCON} = 1 - \frac{\sum\limits_{e^- \text{ is S-covered}} w_s * \mathrm{Typicality}(e^-) + \sum\limits_{e^- \text{ is F-covered}} w_f * \mathrm{Typicality}(e^-) + \sum\limits_{e^- \text{ is D-covered}} w_d * \mathrm{Typicality}(e^-)}{\sum\limits_{e \in \mathrm{NegCov}} \mathrm{Typicality}(e)}$$

where:

PosCov: set of positive events covered by a two-tiered concept description, NegCov: set of negative events covered by a two-tiered concept description,

Typicality(e): the degree of typicality of the event e specified by the expert when the event is given.

w_s: if typicality(e) $\geq t_2$ then 1, else w,

w_f: if $t_2 \geq$ typicality(e) $\geq t_1$ then 1, else w,

w_d: if $t_2 \geq$typicality(e) then 1, else w,

where, $t1$ and $t2$ are thresholds, and $1 \geq t_2 \geq t_1 \geq 0$, $1 \geq w > 0$.

4.3 Accuracy

Now *accuracy* can be defined as a weighted sum of typicality-dependent completeness and typicality-dependent consistency.

Accuracy(description) = w_1*TYCOM(description) +w_2*TYCON(description)

where $w_1 + w_2 = 1$ are accuracy weights, which can be determined by user. One can increase the importance completeness by increasing w_1. The defaults of both $w1$ and $w2$ are 0.5.

4.4 Comprehensibility

As described above, a measure of comprehensibility of a concept description is difficult to define. We will approximate this measure by a computational simplicity, defined as:

$$v_1 \sum_{op \in BCR(dsp)} C(op) \quad + \quad v_2 \sum_{op \in ICI(dsp)} C(op)$$

where:

BCR(dsp): a set of all operator occurrences in the BCR

ICI(dsp): a set of all operator occurrences in the ICI

C(op): the complexity of an operator.

The complexity of operator on the list <interval, internal disjunction, =, <>, not, &, v, implication, predicate> increases with its position on the list. According the order in the list, all operators are assigned a different value as their default values. User can reassign different values to the operators based on the application domain. When an operator is a predicate, C increases with the number of the arguments in the predicate. v_1 and v_2 are comprehensibility weights, $v_1 + v_2 = 1$. The BCR should describe the general and easy-to-define meaning of the concept, while the ICI is mainly used to handle nontypical or exceptional events, therefore the BCR should be easier to comprehend than the ICI. v_1 should therefore be larger than v_2. By default, we assume $v_1 = 0.8$ and $v_2 = 0.2$

4.5 Cost

The cost consists of two parts:

Measurement-Cost -- the cost of measuring the values of variables used in the concept description, defined as the function mc

Evaluation-Cost--the computational cost of evaluating the concept description, defined as the function ec.

$$\mathrm{mc}(\mathrm{description}) = \sum_{e \in Pos+Neg} \sum_{v \in vars(e)} \mathrm{mc}(v)/(|Pos| + |Neg|)$$

$$\mathrm{ec}(\mathrm{description}) = \sum_{e \in Pos+Neg} \mathrm{ec}(e)/(|Pos| + |Neg|)$$

where

vars(e) -- set of all occurrences of variables used to evaluate the concept description or classify the event e.

mc(v) -- the cost of measuring the values of the variable v,

ec(e) -- computational cost of evaluating concept description to classify the event e. This could depend on computation time or on the number of operators involved in the evaluation.

We now define the cost of a description:

$$\mathrm{Cost}(\mathrm{description}) = u_1{*}\mathrm{MC}(\mathrm{description}) + u_2{*}\mathrm{EC}(\mathrm{description})$$

where u_1 and u_2 are cost weights.

With the exception of the weights which can be determined experimentally, we have already defined all three components of the quality measure of concept descriptions: *accuracy*, *comprehensibility* and *cost*. In the next section, we will show how the quality measure evaluates a simple concept description.

5 AN EXAMPLE OF MEASURING QUALITY OF A TWO-TIERED DESCRIPTION

This section provides an example to illustrate the quality measure defined above. The example helps to understand the justification for the chosen criteria, and to compare the results with our intuitive evaluation of the same description.

Description 1

BCR:

(a) *size≠small & (∃x ∃(4)y (seat(x) & leg(y) & ontop(x,y))) V*

(b) *∃x ∃(≥3)y (flat(x) & size(x)=2/3 & leg(y) & ontop(x,y)) V*

(c) *∃x ∃(2)y (seat(x) & wheel(y) & ontop(x,y))*

ICI:

∃x seat(x) & ˘ ∃x backrest(x) => stool

stool => ˘ chair

Description 2

BCR:

∃x ∃(4)y (seat(x) & leg(y) & ontop(x,y))

ICI:

∃x seat(x) & ˘ ∃x backrest(x) => stool

stool => ˘ chair

∃(2)x wheel(x) => Irrelevant(∃(4)y leg(y))

flat(x) & size(x)>2 => seat(x)

Fig. 1 - Two descriptions of the concept"chair" representing different trade-off between the BCR and the ICI.

This example involves measuring the quality of two discriminant descriptions of the concept of "chair", seen as an abstract visual concept. This example is different from the one given in Section 2, since it is based on specific instances of the "chair" concept (see Fig. 2) and is defined in a formal way, as in the INDUCE system (Michalski 80). The instances of visual concepts present a high degree of variability, and are affected by noise and context. For this reason visual concepts can be better represented through a two-tiered scheme, that allows the system to capture the stable characteristics and reason about the special cases in a unified framework.

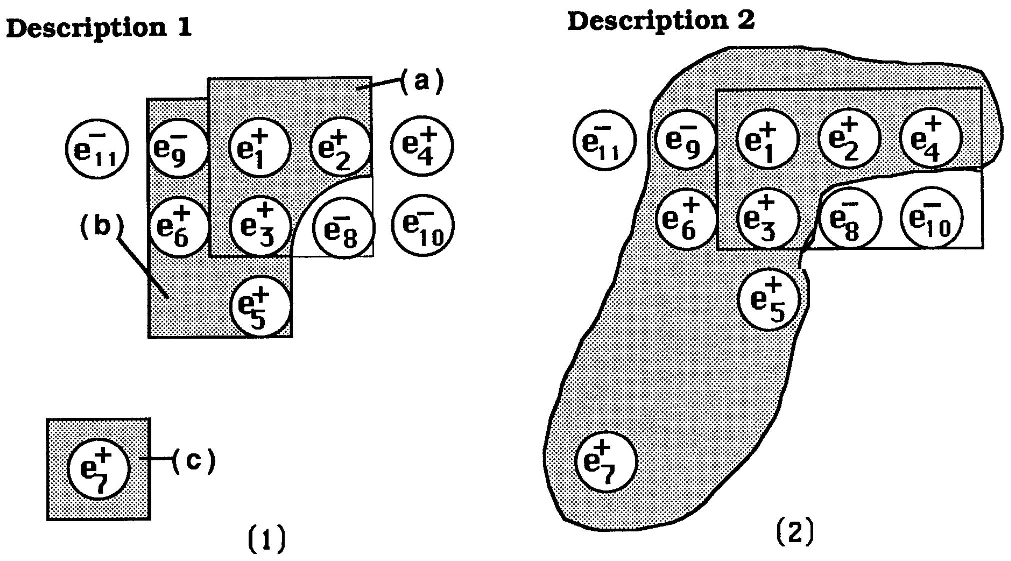
In particular, suppose that we want to evaluate and compare the quality of the two descriptions given in Fig. 1, with respect to the examples given in Fig. 2. Examples e_1-e_7 are instances of the abstract "chair" concept, e_8 and e_{10} are instances of the "stool" concept and examples e9 and e11 are instances of the "sofa" concept. According to the evaluation scheme introduced in the previous sections, we are to evaluate the accuracy of the two descriptions as a first criterion. In order to do this we need to compute the Typicality-dependent Completeness (TYCOM) and the Typicality-dependent Consistency (TYCON). Description 1

Symbol	**Examples**	**Typicality**
e^{+}_{1}	leg(a&b&c,&d) & seat(e) & flat(e) & area(e)=3 & backrest(f)	1.0
e^{+}_{2}	leg(a&b&c&d) & seat(e) & backrest(f)	1.0
e^{+}_{3}	leg(a&b&c&d) & seat(e) & flat(e) & area(e)=2 & backrest(f)	0.9
e^{+}_{4}	leg(a&b&c&d) & seat(e) & size=small & backrest(f)	1.0
e^{+}_{5}	leg(a&b&c) & seat(d) & flat(d) & area(d)=2 & backrest(e)	0.6
e^{+}_{6}	leg(a&b&c&d) & flat(e) & area(e)=3 & backrest(f)	0.8
e^{+}_{7}	wheel(a&b) & seat(c) & backrest(d)	0.4
e^{-}_{8}	leg(a&b&c&d) & seat(e)	0.9
e^{-}_{9}	leg(a&b&c&d) & flat(e) & area(e)=3 & flat(f) & area(f)=3 & backrest(g)	0.9
e^{-}_{10}	leg(a,b,c,d) & seat(e) & size=small	1.0
e^{-}_{11}	seat(a&b) & backrest(c)	1.0

Fig-2 Examples of Chairs

covers positive examples e_1^+, e_2^+, e_3^+, e_5^+, e_6^+, e_7^+ and negative example e_9^-, and description 2 covers positive examples e_1^+, e_2^+, e_3^+, e_4^+, e_5^+, e_6^+, e_7^+ and the negative example e_9^-. The negative example e_8^- could be covered by the BCR part of description 1, but an ICI rule prevents them from being covered by the two-tiered description. This ICI rule says that if an object that would normally be recognized as a chair does not have a backrest, then it is probably a stool, and hence it is not a chair. The same happens for description 2 and events e_8^- and e_{10}^-.

According to the covered examples, and to their typicality, as given in Fig. 2, and if w1=w2=0.5 and w=0.8, the TYCOM measure is 0.81 for the first description and 0.97 for the second description, while the TYCON measure is 0.76 for the first description and 0.81 for the second. The final accuracy measure is then 0.78 for the first description and 0.89 for the second.


Description 1

Description 2

Fig. 3 - The Coverage of Examples in Fig. 2 by Description 1 and Description 2 from fig. 1

This is because the second description is more complete, but also because the most typical events are covered by the BCR, while the non-typical ones are covered through deductive reasoning. The TYCON value of the two descriptions is different, although they cover the same number of negative examples, because the second description does not cover them explicitly.

Moreover, the second description is simpler. The comprehensibility is measured on the basis of the syntactic complexity of the descriptions. The syntactic complexity is evaluated as in the previous section, and is 20.8 for the first description and 18.3 for the second. Comprehensibility would be considered (and measured) by the LEF evaluation scheme only if the tolerance for the accuracy criterion is sufficiently high (higher than 1.1).

Above seems to agree with our intuitive evaluation of the two descriptions, since the second one is shorter and more comprehensible. It covers exceptional examples (such as the wheel chair - example e_7^+) through a reasoning process, rather than by a more complicated explicit description, as the first one does.

6 CONCLUSIONS AND FUTURE WORK

The presented measure of quality of a concept description involves three basic criteria: accuracy, comprehensibility, cost. It takes into account the interrelationships between these criteria in order to capture aggregate characteristics that contribute to quality, but are not measurable without providing more data. Thus, it does not include the predictive power of a description, as discussed in sec. 3, because it is not measurable without more data.

The measure applies to concept descriptions expressed in a two-tiered representation. Generally speaking, it prefers descriptions in which typical events are covered by assertions that are explicit, simple, and efficient to evaluate, and non-typical events are covered through a reasoning process based on the ICI knowledge.

Some experimental results have been obtained using the concept of an "*acceptable labor-management contract.*" The other case examined was the concept of a "chair" (sec. 5). In the experiments, we used the quality measure as a heuristic to search for a better two-tiered concept description starting from a concept description generated by AQ15 or INDUCE, which is discriminant, complete and consistent. The measure was also used to select the final description. The descriptions generated in this way indeed were better than the original ones.

Currently, a larger system that produces and evaluates two-tiered concept representations is being developed. In its current form, the system accepts as input a discriminant, complete and consistent concept description, such as is generated by AQ15 or INDUCE. If the description can be improved, The system produces a two-tiered description of this concept that is qualitatively better. It does so by searching heuristically the space of all two-tiered descriptions. The quality of concept descriptions is the heuristic driving the search. The search operators are generalization and specialization of the description. In its current implementation, generalization is realized by selector truncation, while specialization is realized by complex truncation. The final concept description is selected on the basis of the quality measure.

A number of problems that stem from this work will have to be addressed in the future. First, an integrated system that learns two-tiered concept descriptions from examples needs to be designed and built. Currently, two-tiered descriptions are generated by improving previously learned one-tiered descriptions. The quality of descriptions will then be integrated with the learning algorithm of such a system.

Second, more attention should to be given to technical properties of the selected characteristics of quality. Problems of quality contribution of the implicit part of the description have to be researched in more detail. The question of comprehensibility of a description needs to be investigated through experiments involving human subjects.

REFERENCES

(1) Bergadano, F. , Giordana, A. , Saitta, L., "Automated Concept Acquisition in Noisy Environments", IEEE Transactions on PAMI, July 1988.

(2) Bergadano, F., Giordana, A., "Pattern Classification: An Approximate Reasoning Framework", International Journal of Intelligent Systems, (To appear).

(3) Kemeni, T. G., "The use of Simplicity in Induction", Psychological Review, vol. 62, No. 3, pp. 391-408, 1953.

(4) Michalski, R. S., "A Variable-Valued Logic System as Applied to Picture Description and Recognition", in "Graphic Languages", Nake, F. and Rosenfield, A. (Eds.), North Holland, 1972.

(5) Michalski, R. S., "AQVAL/1--Computer Implementation of a Variable-Valued Logic System VL_1 and Examples of its Application to Pattern Recognition", Proc. of the 1st International Joint Conf. on Pattern Recognition, Washington, D.C., pp. 3-17, 1973.

(6) Michalski, R. S., "Pattern Recognition as Rule-guided Inductive Inference", IEEE Transactions on PAMI, vol. 2, NO. 4, pp. 349-361, 1980.

(7) Michalski, R.S., "A Theory and Methodology of Inductive Learning", Chapter in the book "Machine Learning, an Artificial Intelligence Approach", Michalski, R. S., Carbonell, J. G., Mitchell, T. M. (Eds.), Tioga Pub. Co., Palo Alto, Ca, 1983.

(8) Michalski, R. S., Carbonell, J. G., Mitchell, T. M., "Machine Learning: An Artificial Intelligence Approach", Tioga Publishing Co., Palo Alto, Ca, 1983.

(9) Medin, D. L., Wattenmaker, W. D., Michalski, R. S., "Constraints and Preferences in Inductive Learning: An Experimental Study Comparing Human and Machine Performance", ISG report 86-1, UIUCDCS-F-86-952, Department of Computer Science, University of Illinois, Urbana, February 1986.

(10) Michalski, R. S., Mozetic, Hong, J.,I., Lavrac, "The Multi-purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains", Proc. 5th AAAI, pp. 1041-1045 1986.

(11) Michalski, R. S., "Two-Tiered Concept Meaning, Inferential Matching and Conceptual Cohesiveness" , Chapter in the Book "Similarity and Analogy", S. Vosniadou and A. Ortony, (Eds), 1987.

(12) Mitchell, T. M., "Version Spaces: An Approach to Concept Learning", Ph. D. Dissertation, Stanford University, December 1978.

(13) Mitchell, T. M., "The Need for Biases in Learning Generalizations", Tech. Report, No. CBM-TR-117, Rutgers University, 1980

(14) Pearl, J., "On the Connection between the Complexity and the Credibility of Inferred Models", International Journal of General Systems, vol. 4, pp. 255-264, 1978.

(15) Popper, K., "The Logic of Scientific Discovery", Harper and Row, New York, 1968 (2nd edition).

(16)Sturt, E., "Computerized Construction in Fortran of a Discriminant Function for Categorical Data", Applied Statistics, vol. 30, pp. 213-222, 1981.

(17) Utgoff, P. E., "Machine Learning of Inductive Bias", Kluwer Academic Publ., 1986.

(18) Watanabe, S. , "Knowing and Guessing - a Formal and Quantitative Study", Wiley Pub. Co., 1969.

(19) Weber, S., "A General Concept of Fuzzy Connectives, Negations and Implications based on t-norms and t-conorms", Fuzzy Sets and Systems, vol. 11, pp. 115-134, 1983.