MEASURING QUALITY OF CONCEPT DESCRIPTIONS

by

F. Bergadano
S. Matwin
R. S. Michalski
J. Zhang

Proceedings of the Third European Working Session on Learning, Glasgow, pp. 1-14, October 1988.

PROCEEDINGS OF THE THIRD EUROPEAN WORKING SESSION ON LEARNING

TURING INSTITUTE • GLASGOW

3-5 OCTOBER 1988

Edited by Derek Sleeman

Measuring Quality of Concept Descriptions

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

Artificial Intelligence Center
Department of Computer Science
George Mason University,
Fairfax, VA 22030

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 goes beyond the conventional logic-style descriptions, as it applies to two-tiered concept representations. In such a representation, the first tier characterizes typical and easy to describe properties of the concept, and the second tier describes exceptions, allowed transformations of these properties and changes of meaning of the concept in different contexts. Another novel feature is that the measure takes into consideration the typicality of cases covered by the description.

In the proposed measure, the quality of the descriptions depends on three major criteria: the accuracy, the comprehensibility and the cost. Two exemplary concept descriptions are evaluated to illustrate the method. Presented results are part of a broader research on methods for learning two-tiered concept descriptions.

1 Introduction

Inductive inference is one of the basic strategies of machine learning. Given examples, initial concept descriptions, or any other incomplete concept representation, 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 descriptions one needs a criterion for preferring one assertion over the other.

This paper proposes a general measure for evaluating the quality of concept descriptions. The description quality measure applies to both descriptions supplied by a human, and to descriptions produced by an automated learning system. The measure was specifically designed for evaluating descriptions of flexible and context-dependent 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, in particular, be the completeness and consistency of

---

[1]On leave from University of Torino, Torino, Italy
[2]On leave from University of Ottawa, Ottawa, Canada

generated descriptions with regard to the examples.

Another factor is the predictive power of descriptions, i.e., a measure of description performance on new examples. One may also consider the ease to comprehend descriptions, and to explain them in terms of concepts already known to the system. Finally, one may take into consideration the cost of measuring variables and terms in descriptions, as well as the cost of storing and evaluating descriptions for predicting 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 additional 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 descriptions/hypotheses. To determine the performance on new facts requires that new facts are available. Therefore, such a criterion is not applicable during hypothesis learning, when one needs to choose a hypothesis among various candidates before testing it on new data. This paper is concerned with such a situation, that is, with determining the quality of a description during learning.

The simplicity of a hypothesis has been traditionally the major criterion for evaluating it (e.g., Kemeni, 1953). Popper (Popper,1968) pointed that simpler descriptions are easier to refute, and therefore are preferable. Pearl (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 used in automated learning systems (Michalski. 1980; Bergadano, Giordana, Saitta, 1988).

Broader aspects of the problem of choosing an inductive hypothesis are discussed in (Mitchell, 1980; Michalski, 1983; Utgoff, 1986; Michalski, Carbonell, Mitchell, 1983) Recently, (Medin , Wattenmaker and Michalski, 1986) presented results of psychological testing indicating that humans use not only simplicity but also other criteria for choosing among inductive hypotheses.

The problem of defining the criterion for choosing among assertions generated during learning is a fundamental and still unresolved issue in machine learning. This paper presents a quality measure of descriptions that combines together several factors, specifically, the accuracy, the comprehensibility and the cost.

The accuracy of concept description reflects the degree to which the description relates to the concept it describes. In the case of concept learning from examples, the 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 will increase.

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, it 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, the typicality, the comprehensibility and the cost (for storing and evaluating the descriptions). Second, it can be applied to concepts represented in a two-tiered form.

2 Two-tiered Concept Representation

Before we define the description quality measure, we first describe briefly the basic ideas underlying the two-tiered concept representation. In this representation, any concept is defined by specifying basic features that cover most typical instances of the concept. Groups of such typical instances can be covered by conjuncts. These conjuncts are linked by disjunction. They individually contribute to the accuracy of the description, depending on how many examples they cover or explain. One could think here about an analogy with the Taylor series expansion of a function. The individual terms of this expansion contribute to the precision with which the function is approximated. If a concept description is to include rare or exceptional events, however, other conjuncts, covering only small numbers of events, will have to be added. This may negatively influence the comprehensibility and cost of the description.

To deal with this problem, Michalski (Michalski, 1987) has proposed a two-tiered knowledge representation. A concept description is split into two-parts: a Basic Concept Representation *(BCR)* and an Inferential Concept Interpretation *(ICI)*. The BCR defines the concept explicitly, by giving a description of the concept in terms of either the attributes observed in the example, or in terms constructively learned during concept formation. The prototypical instances of the concept are classified by matching with the BCR. Anomalies, exceptions and context-dependencies are covered by a reasoning process that uses information contained in the ICI. 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 supplied by the BCR (concept *shrinking*). The ICI uses production rules which may be deductively chained to classify "special" uses of concepts. A simple form of ICI may also define a distance measure to classify examples which are close to the existing BCR in the sense of this measure (so called *flexible matching* (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:(no-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 = administrative position) --> (a person in charge of meeting, of a University Department or a court)
(context = prison) --> (specializes to electric_chair)
(a part is broken) --> (a broken chair)

This simple example illustrates several important features of the two-tiered representation. If recognition time is important, only BCR will be used to match an example. If more time can be allocated, or if a more precise classification is required for a given event, ICI is used. ICI relies on background and general knowledge, and on the context in which the concept operates. Contexts can have hierarchical organization. Finally, ICI rules may chain, although it is not shown in this simple example.

We argued 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 was improved, since the two-tiered description is more complete and consistency has not changed. Second, comprehensibility has increased, since the prototypical properties of the chair concept are separated from its possible modifications and specializations.

Some systems that generate and use two-tiered representations have been described in the literature (Michalski et al., 1986, Bergadano, Giordana, Saitta, 1988, Bergadano, Giordana, [to appear]). Two-tiered concept descriptions are usually simpler, easier to understand and more efficient to use than the conventional ones. They also exhibit performance improvement on a testing set. In the systems developed so far, the ICI is based exclusively on flexible matching. More importantly, in their quality evaluation measures, these systems do not take into account the inferentially covered parts of concept descriptions. Improvement in quality is therefore measured only by the improvement in the first tier.

As with any measure of concept description quality, the final evaluation is only possible with the use of a testing set. However, for the reasons that we have already discussed, it is necessary to evaluate during learning the quality of the inductive hypotheses. This paper defines the quality of a concept description in a general way. General requirements for such a measure are specified, and a specific measure is defined.

3 Criteria for Determining the 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, as well as a mechanism for combining them into a a single measure.

• The *accuracy* represents the description's ability to produce correct classifications. A common way to prefer more accurate descriptions is to require that they be complete and consistent 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 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 - Applied Statistics, 1981).

Even if a description is incomplete and inconsistent, the number of positive and negative examples it covers provides important information for evaluating its quality. In this case, we can measure the degree of completeness and consistency of a given description. 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, on an inductive basis. Therefore, one can be confident in the information contained in the BCR only if a sufficient number of examples is available, or if the examples are really typical or representative for the domain. On the contrary, the ICI, being generated by experts or with the available domain knowledge, is appropriate when dealing with 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 from deduction to induction (Michalski, 1987). These requirements are met by making 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 typicality of examples at the time they are presented to the system responsible for building the initial description.

The degree of generality is also related to accuracy, since it affects prediction power. Given the same completeness and consistency, a learning system should prefer maximally specific characteristic descriptions and 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. 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.

Inductive learning systems that are only based on objective evaluation criteria, such as statistical inference, have been developed in the past. Recent research in this field, however, emphasizes knowledge intensive approaches. In this case, assertions are evaluated also on the basis of subjective or extra-evidential considerations, in order to explicitly take into account the available domain knowledge. In inductive learning systems, domain 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.

• The *comprehensibility* of the acquired knowledge is related to subjective and domain dependent criteria. An important requirement of an AI system is that knowledge has to be explicit and easily understandable by human experts. This is important for improving or modifying the knowledge, and for communicating with experts. A black box classifier will not be accepted by experts as a help in their work, therefore knowledge acquired automatically should be easy to understand, should contain the descriptors most frequently used by experts, and should not be syntactically too complex. In practice, only the last feature is easy to obtain.

• The *cost* captures the properties of a description related to its storage and use. 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.

These 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 formula that, given the numeric evaluations of single criteria, produces a number that represents their combined value. Examples of such approaches are multiplication, weighted sum, maximum/minimum, t-norm/t-conorm (Weber, 1983). Although these approaches are often appropriate, some of them may present disadvantages. Firstly, they usually combine a set of heterogeneous evaluations into a single number, and the meaning of this final number is hard to understand 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.

In order to overcome some of these problems, we use a *lexicografic evaluation functional* (LEF) (Michalski, 1972, Michalski, 1983) that combines the above mentioned criteria. The general description quality measure is thus defined as:

$$GDQ(description) = \langle(Accuracy,\tau_1), (Comprehensibility,\tau_2), (Cost,\tau_3)\rangle$$

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 is associated with each criterion. If the difference of the evaluation of two expressions under a given criterion is less than the corresponding tolerance, the two descriptions are considered equivalent with respect to that criterion. The most important measure in the LEF is evaluated first, and the subsequent measure is evaluated only if the previous one is a tie.

The LEF evaluation scheme is not affected by the main problems which affect algebraic formulas which we have discussed above, but it may still be useful to extend it in some cases. 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 is very 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 above criteria can also be applied to two-tiered descriptions. The accuracy of the acquired knowledge does not only depend on the explicit information, but also on the 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:

Quality(description) = <(Accuracy, τ_1) (Comprehensibility, τ_2) (Cost, τ_3)>

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

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

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, 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 are made mutually exclusive: if an event is S-covered, then it is not D-covered or F-covered, and if an event is D-covered, then it is not F-covered. Thus, S-covered events are explicitly covered, and F-covered and D-covered events are implicitly covered. In general, descriptions that cover many typical positive events are most preferred. Completeness is therefore proportional to the typicality of the events covered. Moreover, if negative events are covered, the consistency of the description is inversely proportional to the typicality of the negative events covered.

It is also preferred that the typical events are covered by the BCR , and non-typical, or exceptional events are 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 ICI, on the contrary, is deductively obtained from the background knowledge, or from a human expert and relies more on general and domain knowledge, and it is more reliable when dealing with the special or rare cases. For these reasons, a typical positive explicitly-covered event should contribute to completeness more than implicitly-covered. And vice-versa, a nontypical positive implicitly-covered events contribute to completeness more than 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. A similar argument holds for consistency.

Now, we define the typicality-dependent completeness (TCOM) and typicality-dependent consistency (TCON) of a description:

$$\mathrm{TCOM} = \frac{\sum_{e^+ \text{ is S-covered}} w_S * \mathrm{Typicality}(e^+) + \sum_{e^+ \text{ is F-covered}} w_F * \mathrm{Typicality}(e^+) + \sum_{e^+ \text{ is D-covered}} w_D * \mathrm{Typicality}(e^+)}{\sum_{e \in \mathrm{PosCov}} \mathrm{Typicality}(e)}$$

$$\mathrm{TCON} = 1 - \frac{\sum_{e^- \text{ is S-covered}} w_S * \mathrm{Typicality}(e^-) + \sum_{e^- \text{ is F-covered}} w_F * \mathrm{Typicality}(e^-) + \sum_{e^- \text{ is D-covered}} w_D * \mathrm{Typicality}(e^-)}{\sum_{e \in \mathrm{NegCov}} \mathrm{Typicality}(e)}$$

where:

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

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

typicality(e): 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, t_1 and t_2 are thresholds, and $1 \geq t_2 \geq t_1 \geq 0$, $1 \geq w > 0$.

Now *accuracy* can be defined in terms of TCOM and TCON:

Accuracy(description) =
w_1*TCOM(description)
+w_2*TCON(description)

where $w_1 + w_2 = 1$.

A measure of comprehensibility of a concept description is difficult to define. We will approximate this measure by a syntactic complexity, defined as:

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

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. When an operator is a predicate, C increases with the number of the arguments in the predicate.

v_1 and v_2 are 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.

The cost consists of two parts:

Measure-Cost -- the cost of measuring the values of variables used in the concept description, it is defined as the function MC

Evaluation-Cost-- the computational cost of evaluating the concept description, it is defined as the function EC.

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

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

where

vars(e) -- set of all occurrence variables used to evaluate the concept description to 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:

Cost(description) = u_1*MC(description) +
u_2*EC(description)

where u_1 and u_2 are 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. In the next section, we will show how the quality measure evaluates a simple concept description. This quality measure has been experimented with two non-trivial examples, acceptable labor-management contract and the concept of "chair", the results are satisfactory. Currently, we are

implementing the quality measure in a two-tiered concept learning system and using it to guide the search for a better two-tiered 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.

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 modifications related to 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 e1-e7 are instances of the abstract "chair" concept, e8 and e10 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 (TCOM) and the Typicality-dependent Consistency (TCON). Description 1 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}^-.

BCR:

size≠small & (∃x ∃(4)y (seat(x) & leg(y) & ontop(x,y)))
∃x ∃(≥3)y (flat(x) & size(x)=2∨3 & leg(y) & ontop(x,y))
∃x ∃(2)y (seat(x) & wheel(y) & ontop(x,y))
(1)

ICI:

∃x seat(x) & ⌉∃x backrest(x) => stool
stool => ⌉chair

BCR:

∃x ∃(4)y (seat(x) & leg(y) & ontop(x,y))
(2)

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.

		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)	1.0
e_4^+	leg(a&b&c&d) & seat(e) & size=small & backrest(f)	0.9
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 abstract visual concepts

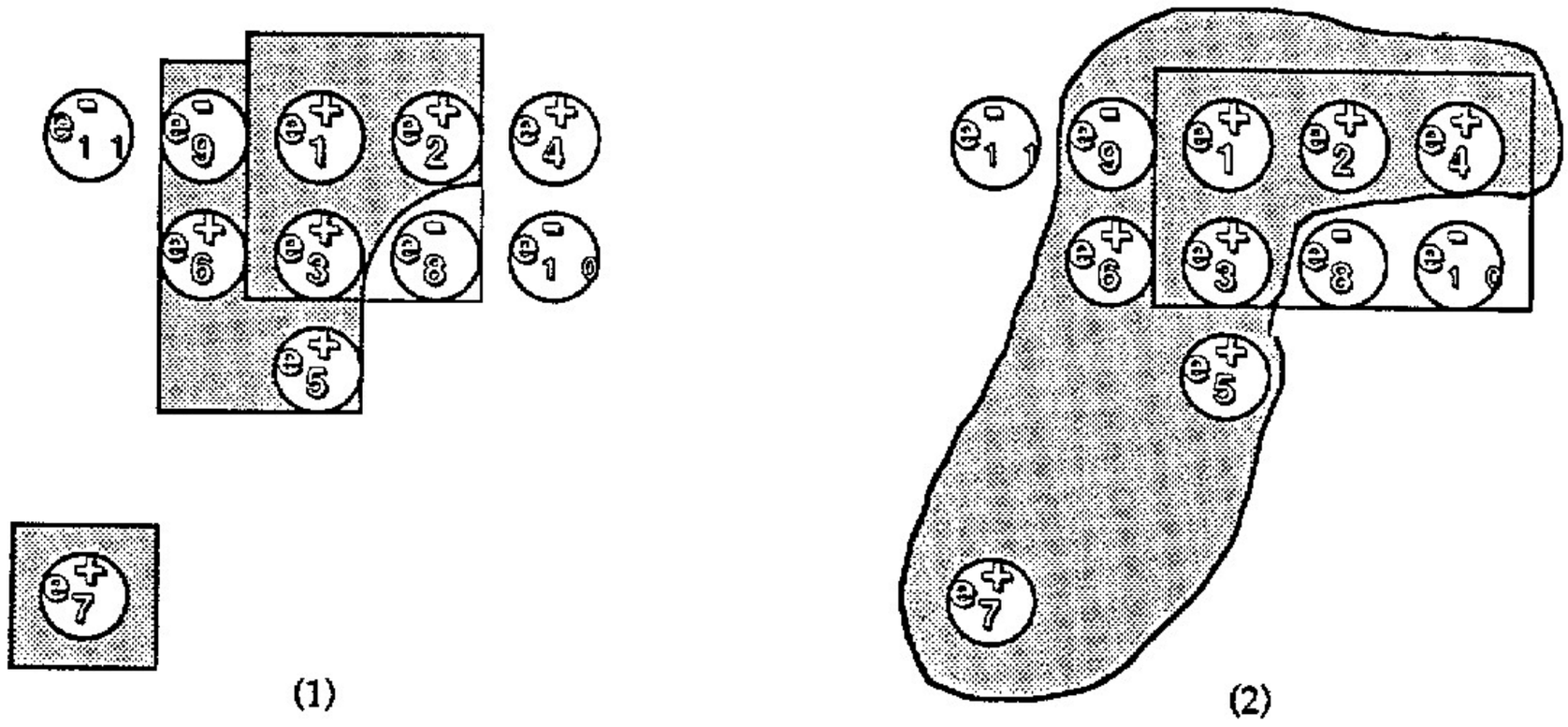

Fig. 3 - Coverage of the two descriptions in Fig. 1 for the examples in Fig. 2

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 TCOM measure is 0.81 for the first description and 0.97 for the second description, while the TCON 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. 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 TCON 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).

This 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, the accuracy, the comprehensibility and the cost. It takes into account the interrelationships between these criteria in order to capture more aggregate characteristics that contribute to the quality, but are not directly measurable. Predictive power of the description, as discussed in sec. 3, is an example of such a characteristic.

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 discriminant, complete and consistent concept description generated by AQ15 or INDUCE. 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 generated by AQ15 or INDUCE. The system produces a two-tiered description of this concept that is qualitatively better, if there is one. It

does so by searching heuristically the space of all two-tiered descriptions. The quality of concept descriptions is the heuristic driving the search. 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 than 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.

Acknowledgements

This research was done in the Artificial Intelligence Center of George Mason University. The activities of the Center are supported in part by the Defence Advanced Research Projects Agency under grant, administered by the Office of Naval Research, N00014-85-k-0878, and in part by the Office of Naval Research under grant No. N00014-88-K-0226. 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 number A2480. The authors thank Joan Elliott and Ken Kauffman for the comments.

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", Stella Vosniadou and A. Orton, (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, 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.