An Integrated Characterization and Discrimination Scheme to Improve Learning Efficiency in Large Data Sets

Roberto GEMELLO, Franco MANA

CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.a.
via G. Reiss Romoli, 274 - 10148 Torino (Italy)

Abstract

This work proposes a learning scheme which integrates Characterization and Discrimination activities with the aim of improving learning efficiency in large data sets. Characterization is considered to be a process which builds up a rough concept description using only positive examples. This description already excludes most of the extreme negative examples. Discrimination is considered to be an incremental learning process which begins with the characteristic description and refines it so as to make it consistent with the negative examples (near misses) which are still covered. During this phase learning efficiency is greatly improved by considering only near misses as counter-examples. Finally, the description is simplified by dropping some characterizing but not discriminant parts of the description.

This learning scheme is discussed and compared with the traditional data reduction techniques. Some experimental results are reported which show the gain in efficiency obtained, particularly on real applicative domains.

1 Introduction

Machine Learning has been one of the main theoretical topics of AI research over the past decade and several learning techniques have been widely investigated. Among them Inductive Learning from Examples has reached advanced level and is beginning to move out of the labs and face real applicative problems. In taking this step the learning systems must change their techniques of dealing with the few hand coded examples of artificial domains and they must be adapted so that they can manage the large number of real data present in the environment in which the learning system is operating. On one hand this impact with large data bases of samples is positive for the learning systems: in fact they can demonstrate their ability to learn rules which are not merely a summary of the examples but incorporate an ability to make predictions, which, in turn, can be tested on statistically relevant test sets. On the other hand, an increased number of samples can cause inefficiency due to the complex computations involved, especially in the systems which use a first order representation language, and their greater learning power calls for much greater computational effort.

For this reason it is necessary to study techniques which will allow the examples to be used more efficiently in the inductive process. These techniques have often been called Data Reduction Techniques [Michalski and Larson, 1978, Cramm, 1983, Pollack, 1983] and their aim is to cut down computational effort by reducing the number of examples involved in the learning process, without compromising the meaningfulness of the learned knowledge. Unfortunately, the methods which have been proposed up to now are either inadequate for a first order representation language or, in their turn, arc computationally too expensive. After a brief revision of the present state of the art, a new data reduction technique is presented, which adopts an approximation of the characterization as the evaluation criterion to select the counter-examples for each class. This technique is a return to the classical concept of near miss introduced by Winston [1979], and proposes a more operational definition of the same concept, that is used to reduce the number of counter-examples which have to be taken into account during the discrimination process.

The main idea is as follows: first a costless approximation (p* of the characterization (p is computed for each class, using only the positive examples; then, the class counter-examples covered by cp' are defined as near misses of the class (the most useful counter-examples for the computation of a discriminant description of the class); finally, a discriminant description y of the class is obtained through an incremental learning process which, taking cp' as the starting hypothesis, specializes and simplifies it so as to make it consistent and less complex. The learned knowledge (a first order discriminant formula for each class) is proved to be complete and consistent (within a prescribed tolerance) with the original examples and counter-examples. In fact examples are not arbitrarily dropped but they are used more rationally so that there is no loss of information.

In the technique proposed the characterization process can be seen as a bootstrap procedure which provides the discrimination process with a rough hypothesis which has to be refined using the counter-examples which are still covered. From another point of view this technique constitutes a proposal which could be applied to the problem of integrating characterization and discrimination activities so as to form a learning scheme for conceptual discrimination.

2 Data Reduction Techniques

Data Reduction techniques are embedded in systems which learn from examples in order to take into account the decline in performance when large data sets are considered. This decline is especially relevant in learning systems which use a first order representation language: in fact, this kind of languages offers greater learning power (rules to discriminate structured concepts may be learned) but it has to be paid for by an increased computational effort. A data reduction method aims at cutting down the computational effort by reducing the number of examples which have to be considered during the learning process.

In the following sections some techniques which have been proposed in the literature are briefly outlined. They can be grouped in two different approaches: reduction using evaluation criteria and reduction using compression techniques. Finally, we propose a new method which performs a data reduction by means of an innovative evaluation criterion.

2.1 Using Evaluation Criteria

Given a set of examples E, the reduction process of E consists of finding a new set E' that satisfies the condition:

$$E \supset E' \qquad (2.1)$$

The reduced set E' is obtained by selecting a subset of elements which belongs to E (2.1). The elements which belong to the set E' are chosen using an evaluation criterion: an example which belongs to E is an element of the set E' if the evaluation criterion suggests it. The aim of the evaluation criterion is to select the best examples to solve the learning problem. A possible definition for the evaluation criterion, proposed by Michalski in [Michalski and Larson, 1978, Cramm, 1983], is aimed at selecting the best examples to accomplish the discrimination task: it gives an estimate of the distance between two examples, i.e. it determines how similar two different examples are. The distance is measured between an example belonging to the set E and a fixed example used as a reference point. Then, the examples of the new data set are chosen following a strategy which selects the examples that represent border line concepts.

The main advantage of the evaluation criterion approach is that a simple evaluation criterion can be defined which ensures a high level of efficiency. On the other hand, the definition of the evaluation criterion is a difficult problem, because in the new data set some examples are dropped and only a good evaluation criterion preserves the effectiveness of the knowledge acquired during the subsequent learning process. Further, the evaluation criteria proposed in the literature, such as the distance measure proposed in [Michalski and Larson, 1978], are suitable for examples expressed in an attribute-value language, but are not applicable to examples expressed in a first order language.

In Figl.a the typical behavior of an evaluation criterion based on a data reduction technique is reported.

2.2 Using Compression Techniques

Given a set of examples E, the reduction process of E consists of finding a new set E' that satisfies the following conditions:

$$|E'| < |E| \qquad (2.2)$$

$$\forall c \in E, \exists e' \in E' \quad c' |> e.$$

$$\forall e' \in E', \exists e \in E \quad e' |> e. \qquad (23)$$

where |> means *more general than.*

The cardinality condition imposed by (2.2) is obtained by creating a set of new examples which generalize the original ones (2.3). A new example e'e E' is the most specific generalization of a small set of examples belonging to E. In others words, each subset of E containing examples which are very similar can be compressed (generalized) into a new example. There are various ways of performing the compression, which arc more or less complex, depending on the generalization rules involved. The simplest form of data compression introduces the disjunction connective in the data representation language. In this way, data compression is performed by grouping the similar examples into disjunctive descriptions |Frediani and Saitta, 1987]. An interesting form of data compression involves the dropping condition rule. With this method the data are analyzed in order to identify the features which are irrelevant to the solution of the learning problem in the specific application domain. After that, the data compression is performed by dropping the irrelevant features from the examples. Some experiments in this direction are reported in [Pollack, 1983]. Finally, the more complex form of data compression requires a new representation language in which more general descriptions can be expressed [Frediani and Saitta, 1987]. In this way, data compression is performed by rewriting the input examples in terms of the new language. This is a difficult form of compression because it involves a constructive learning mechanism.

The main advantage of the compression approach is that no information is lost during the compression process. In fact, the new examples generalize the original ones and so any concept description which generalizes the new examples will also generalize the original ones.

The main problem is that the new examples generated by the compression process must not be overgeneralizations of the original ones. In others words any new example created by the compression of a set of original examples (instances of the class H_i) must preserve consistency, i.e. it must not be a generalization for some example of another (disjoint) class. In the general case such a process turns out to be in its turn a learning activity, with a computational complexity which leads to no significant gain in efficiency in the global learning process (data reduction + inductive learning).

In Fig 1 .b the behavior of a compression based data reduction technique is reported.

2.3 Using Concept Characterization as Evaluation Criterion

Our proposal of data reduction follows the evaluation criterion approach. We argue that performing data reduction with an evaluation criterion approach is more suitable in real domains, where large data sets are to be managed. In fact, a good (and simple) evaluation criterion may dramatically reduce the number of examples to be used in the more time consuming phases of the inductive process, without itself being a complex operation. On the contrary, compression techniques may have the same complexity as the learning process, especially when a first order representation language is used.

The crucial point in this approach is the definition of the evaluation criterion. Let's build up a conceptual discrimination framework where there is a set $H=\{H_1, \ldots, H_n)$ of conceptual classes to be learned and for each class a set E(Hi) of examples of the class. Given a class Hj, E(Hj) is the set *of positive examples* of Hj while $CE(H_i)$ = $uj{=}_iE(Hj)$ is the set of *negative examples* or *counter-examples.* The evaluation criterion must establish which subset of examples is to be used and in which phase of the learning process they must be used. The central idea is to use (an approximation of) the characterization of each class Hi as the criterion to select the examples and counter-examples which are to be used during the discrimination phase. The perfect characterization Φ of a class is a description that states all the facts and relations that are true of all examples in the class. An approximation Φ' of the characterization is a generalization of Φ so that EXT(Φ') 3 EXT(Φ), where EXT(Φ is the extension of the description £ (the set of all the examples described by Q.

Given a class Hj, the proposed evaluation criterion to reduce the input data necessary to discriminate the class H_i from each other class H_j (j not= i) can be split in two parts.

For positive examples:

1. use all the positive examples to compute an approximation <p'(Hj) of the characterization $\Phi(H_i)$;
2. use the subset E'(Hj) = $E(H_j)$ n EXT(Φ(Hi)) during the discrimination phase.

For negative examples:

1. use no negative examples during the characterization;
2. use the subset $CE'(H_i) = CE(H_i)$ n $EXT(\Phi'(H_i))$ during the discrimination phase.

The criterion for positive examples is aimed at removing the noisy examples from the set E(Hj) of positive examples of the class Hj; in fact if an example does not belong to EXT(Φ'(Hi)) because does not make allowance for some fact or relation common to most of the examples of the class, then it can be considered as a spurious example. The criterion for negative examples is the formalization of the intuitive idea that some counter-examples are more useful than others when learning discriminant descriptions. This idea was first introduced by Winston with the near miss concept [Winston, 1979]. Near misses are the most useful counter-examples with which to learn a discriminant description of a class, because they avoid overgeneralizations and ensure that the description is specified by the facts and relations which are necessary to discriminate the class. Winston defined a near miss as *"a sample which does not qualify as an instance of the class being taught for some small number of reasons"* [Winston, 1979, p.32]. Such a definition is not operational, because the description of the class being taught is not known in advance and so we do not know how to select the near misses. We present an operational version of the above definition as follows: *"a near miss is a counter-example which belongs to the extension of an approximation of the characterization of the class being taught".* The operationalization consists in using the approximation of the characterization instead of the (unknown) class definition. The counter-examples which do not qualify as an instance of the class for "some small number of reasons" are those which are covered[1] by the approximation of the characterization. They are selected as near-misses: in fact they are negative examples and are quite similar to most of the positive examples of the class. By adopting this technique of selecting the near misses we can consider the characterization to be a process which also performs a rough discrimination and so provides the discrimination process with an initial hypothesis and selects the useful counter-examples. This initial hypothesis is then specified to make it consistent with regards to near misses, and is simplified in order to reduce its complexity, by means of the incremental learning discrimination process.

an example e is covered by a description £ if eE EXT(Q.

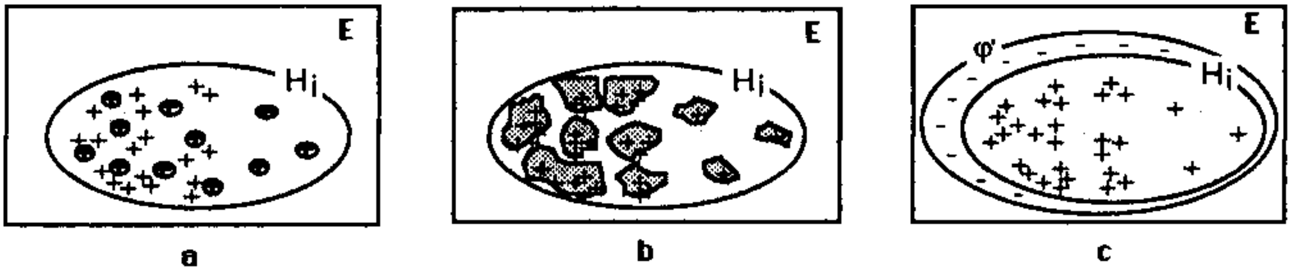

Fig 1 - Data reduction can be performed following two different approaches: using an evaluation criterion or performing data compression. In the first approach (a) the data reduction is achieved by means of the selection of the best concept examples (circled +). In the second approach (b) the reduced data set is composed of new examples (shaded shapes), which are generalizations of the old ones. In the picture (c) a specific evaluation criterion is presented to generate the best data for the concept discrimination: for each class H- a concept approximation Φ' is used to select the concept near misses.

The main advantage of this data reduction proposal is that there is no loss of information. In fact no example is arbitrarily dropped but they are more efficiently used.

A disadvantage is that, for very simple domains with well separated classes, the effort needed to learn the approximation of the characterization can be greater than the effort of the discrimination alone. As a consequence this method appears to be especially well suited for real domains with classes, which are difficult to separate.

In Fig 1 .c the desired behavior of the data reduction analysis using a characterization module as evaluation function is reported.

3 Integration of Characterization and Discrimination Modules

In the previous sections we have suggested a new technique which makes use of the characterization process as an evaluation criteria in order to reduce the input data for the discrimination process. From another point of view this technique can be considered to be a model for the integration of the characterization and discrimination modules of an inductive learning system aimed at solving more complex discrimination tasks.

The proposed learning scheme integrates characterization and discrimination modules with the aim of improving the efficiency of the learning process and the robustness of the learned knowledge. The input of the process for each conceptual class $H_i \in \mathcal{H}$, is a set of examples of the class E(Hj) while the output for each conceptual class $H_i \in \mathcal{H}$, is a discriminant rule: $\psi::>H_i$[2], which is a compromise between a purely sufficient condition and a necessary condition. The learning process is divided into three sequential steps.

Step 1 - Characterization: The positive examples of Hj are analyzed by the characterization module in order to find an approximation $\varphi'(H_j)$ of the concept characterization. During this step learning is performed from only positive examples. Learning the most specific characterization is a computationally very hard task, but, for our aims, it is necessary for this module only to find an approximation of the most specific description.

Step 2 - Discrimination: The concept approximation $\varphi'(H_i)$ is used to compute the concept's best positive and negative examples (near misses) and these new data sets arc the input of the discrimination module. Further, the discrimination analysis uses concept approximation as an initial hypothesis from which to start the inductive Process. The result of this step is a discriminant rule $\rho::>H_i$. This step performs incremental learning using a reasonable concept definition ($\varphi'(H_i)$) as the initial hypothesis and the near misses of the concept as elements to specialize this hypothesis.

Step 3 - Simplification: The output of the discrimination is a description ρ which contains all the necessary conditions inherited by the approximation of the characterization $\varphi'(H_i)$. At this point a simplification is used in order to reduce the complexity of the final knowledge $\psi::>H_i$. The rule simplification can be defined as follows: given a rule $\rho::>H_i$, the set of positive examples of H_i covered by ρ, namely $E(\rho,H_i)= E(H_i)\cap EXT(\rho)$, and the set of the negative examples covered by ρ, namely $CE(\rho,H_i)=CE(H_i)\cap EXT(\rho)$, then the new rule $\psi::>H_i$ is a simplification of $\rho::>H_i$ if and only if the following conditions are satisfied:

$$\mathrm{Complexity}(\psi) < \mathrm{Complexity}(\rho), \tag{4.1}$$

$$E(\psi,H_i) \supseteq E(\rho,H_i), \tag{4.2}$$

$$CE(\psi,H_i) = CE(\rho,H_i). \tag{4.3}$$

As the complexity of a formula is defined as the number of language terms involved in the formula, the simplification process can be viewed as a generalization task which is performed by using the dropping condition rule. Condition (4.3) is a constraint which is very difficult to satisfy because the negative examples $CE(H_i)-CE(\rho,H_i)$ must be considered and this involves a considerable computational effort. However, a realistic simplification process can perform only partial simplification. A possible

[2] $\xi::>H_i$ means that if an unknown example is covered by E, the example is an instance of H_i.

strategy, with which partial simplification is achieved, considers the description ρ as a conjunction of closed subformulae $\rho = \rho 1 \wedge \rho 2 \wedge \ldots\ldots \wedge \rho_n$. In such a case $EXT(\rho)=\cap_i EXT(\rho_i)$ and ψ can be built simply by dropping the ρ_i which are not useful for constraining the $CE(\rho,H_i)$. As $EXT(\rho_i)$ is known for all ρ_i from the previous steps (1,2), ψ may be obtained unexpensively by looking for the simpliest conjunction of ρ_i terms which satisfy (4.3), e.g. with a minimum cost search.

4 Experimental Results

In this section we are going to discuss some experimental results in order to verify the soundness of the learning scheme which has been proposed. The experiments have been carried out in order to compare the computational performance of the learning system in two configurations: in the first (the traditional configuration) the discrimination module is used in isolation; in the second (the learning scheme proposed here) both characterization and discrimination modules are used in an integrated way, with the characterization being used as an evaluation criterion in order to reduce the number of examples for discrimination. The experiments are case studies of learning discriminant rules from examples.

Two different application domains are introduced and the time needed to discriminate each class is reported. The results described below have been obtained using an initial version of the proposed learning scheme which has been implemented on top of the RIGEL inductive learning tool [Gcmello *et al.,* 1988], with a TI Explorer hardware. The characterization module parameters have been tuned so as to create a very simple concept approximation consisting in a conjunction of numerically quantified formulae involving a single object (3(x)N y(x)) (where N can assume the following forms (= n), (> n), (< n) and (e [n..m j) and *y* is a first order formula).

The first experiment concerns the trains domain, a well known artificial domain introduced by Michalski [6j. In Fig.2 the input examples are shown. The problem is to discriminate the trains going east from the trains going west. Each train is described in a first order language. The domain features which have been chosen for the discrimination problem are the same as those proposed in [6]. A carriage is characterized by its shape, number of wheels, length and number of loads while a load is characterized by its shape. Two binary relations are used: the INFRONT relation to describe the carriages sequence and the CONT-LOAD relation to specify which loads are contained into a carriage.

Using only the discrimination module, the learning strategy implements an inductive inference which takes into account all the examples and counter-examples. The system finds complete and consistent descriptions for both the classes and in particular:

∃(x) (length(x) = short)
(car-shape(x) = closed_top) ::> east
(a short and closed top carriage exists) time = 13 sec.

∃(x) (position-infront(x) = 3)
(car-shape(x) = (open-top jagged-top))
(∃(y)(= 2) (length(y) = long)) ::> west
(there exist two long carriages and the third carriage is open or jagged top) time = 15 sec.

Using characterization and discrimination modules in the integrated mode, the system establishes the following discriminating rules:

∃(x) (∃(y)(∈ [2..4]) (#loads(y) = 1))
(∃(y)(∈ [1..2]) (load-shape(y) = triangle))
(car-shape(x) ≠ (jagged-top u-shape)) ::> east
(there exist between two and four carriages with one load,
there exist one or two triangular loads and
a carriage with shape different from
jagged top or u-shape exists) time = 6 sec.

∃(x) (position-infront(x) = 3)(car-shape(x) = (open-top jagged-top))(∃(y)(= 2) (length(y) = long)) ::> west
(there exist two long carriages and the third carriage is open or iagged ton) time = 13 sec.

For the trains going east, the near misses set contains only the west train ev7 which has 2 or 4 carriages with one load and 1 or 2 loads with a triangle shape. For the trains going west the near misses are the trains ev3 and ev5. In fact, they arc the trains with exactly 2 long carriages (the engine is considered a long carriage).

The second experiment is set in a 2D image recognition domain. The results which have been obtained are more interesting in this case because image recognition is in a real applicative domain where noise affects the images and a larger learning set is available. The goal consists in the discrimination of printed capital letters. The input data are generated from a set of 2D images of the letters, produced by a TV camera. The pixel map is analyzed by a low level module which describes the contour in terms of three primitives: angle, straight line and curve. For each primitive a set of features, which are independent of rotation and translation movements, are captured. A binary relation NEXT is defined in order to specify the primitive sequence in the contour. The experiment has been carried out starting from 198 input examples distributed over 18 letters. In the Table I the performance of the system in both the configurations are summarized. The reduction factor of the computational effort when the integrated approach is used is, in this domain,= 3.3.

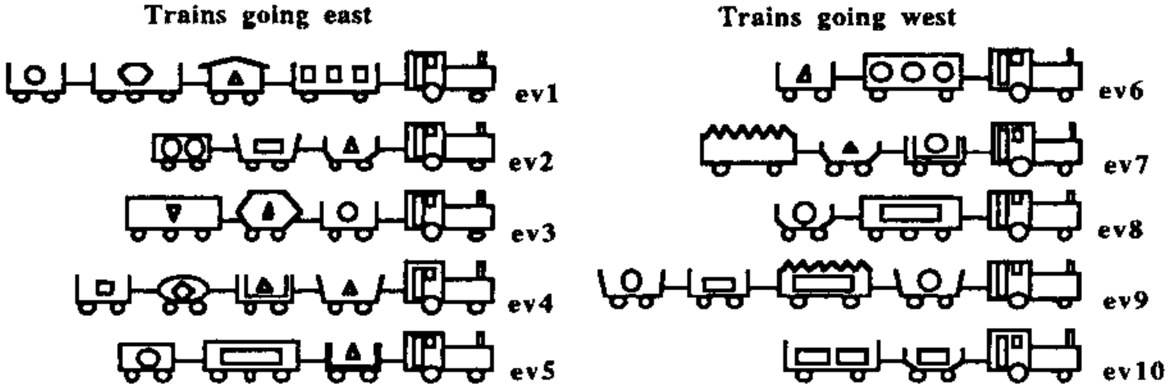

Fig. 2 - The input examples for the trains domain (from Michalski).

Learning Strategy	Discrimination Only	Characterization and Discrimination	Time Ratio
Average Time	1016 sec.	311 sec.	3.3

Table I - Computational performances in the 2D image recognition domain.

5 Conclusions

Inductive learning systems which use a first order representation language pay for their increased learning power in terms of computational complexity. There are two ways of controlling this complexity which are commonly used: heuristic methods are used to guide the inductive search and data reduction methods are adopted to reduce the number of examples which have to be considered.

This paper has focused attention on data reduction methods and has proposed a new technique which exploits the results of a preliminary characterization analysis so that the number of examples and counter-examples to be taken into account during the discrimination analysis can be reduced. From a theoretical point of view it has been shown that the method docs not suffer from the danger of loss of information, as many data reduction methods do. In fact there is not a preliminary trimming of the examples, but instead, thanks to the integration of the characterization and discrimination activities, they are used more rationally. From an experimental point of view, the method which has been proposed has proved to be more suitable in real domains than in simple artificial domains. In fact: (1) it is not guaranteed that the simplest discriminant description will be found but that a discriminant solution with some characterizing elements which is more robust to misclassifications will be provided. This is not of interest in toy domains, but it is essential in real applications; (2) if the data set is small, the characterization complexity may offset the gain in efficiency which is obtained during the discrimination phase, so that there is no improvement in efficiency.

This has been tested in two domains and while in the artificial *trains domain* the method proposed worked well, providing a modest gain (1.4), there was a significantly more relevant gain (3.3) when it was used in the real domain *of 2D image recognition* .

Acknowledgements

The authors would like to thank Lorenza Saitta for helpful comments on drafts of this paper.

References

[Cramm, 1983] S.A. Cramm: "ESEL/2". Report No. 901. Dcpartement of Computer Science University of Illinois at Urbana-Chaimpaign, Urbana, January 1983.

[Frediani and Saitta, 1987] S. Frediani and L. Saitta: "Knowledge Base Organization in Expert Systems". Lecture notes in Computer Science, Vol 286, pag. 217-224, 1987.

[Gemello *et al..,* 1988] R. Gemello, F. Mana and G. Viano: "Inducing Conceptual Discrimination Rules from Examples: an Application to Image Recognition". International Simposium on Methodologies of Intelligent Systems, ISMIS 88, Turin, October 1988.

[Michalski, 1980J R.S. Michalski: "Pattern Recognition as Rule-Guided Inductive Inference". IEEE Transactions on pattern analysis and Machine Intelligence, Vol. PAMI-2, (1980).

[Michalski and Larson, 1978] R.S, Michalski and J.B. Larson: "Selection of Most Representative Training Examples and Incremental Generation of VL| Hypotheses". Report No. 867. Departement of Computer science University of Illinois at Urbana-Champaign, Urbana, May 1978.

[Pollack, 1983] J. Pollack: "Relavant Variable Selection for Inductive Learning Programs". Report No. 902. Departement of Computer Science University of Illinois at Urbana-Chaimpaign, Urbana, January 1983.

[Winston, 1979] P.H. Winston: "Artificial Intelligence". Addison-Weslcy Publishing Company, Menlo Park, April 1979.