Reports

Machine Learning and Inference Laboratory

Generating Alternative Hypotheses in AQ Learning

Ryszard S. Michalski

MLI 04-6
P 04-8
December, 2004

School of Computational Sciences

George Mason University

GENERATING ALTERNATIVE HYPOTHESES IN AQ LEARNING

Ryszard S. Michalski
Machine Learning and Inference Laboratory
George Mason University
and

Institute of Computer Science
Polish Academy of Sciences

michalski@mli.gmu.edu

Abstract

In many areas of application of machine learning and data mining, it is desirable to generate alternative inductive hypotheses from the given data. The A^q-ALT or, briefly, ALT method, presented in this paper, generates alternative hypotheses in two phases. The first phase proceeds according to the standard A^q algorithm, but each star generation process produces not just one *best complex,* but rather a collection of complexes, called *the elite*. This phase ends when the union of best complexes constitutes a complete and consistent cover of the target set, called the *primary hypothesis.* The second phase derives alternative hypotheses by multiplying out the disjunctions of symbols representing complexes in each elite, and creating an irredundant DNF expression. Individual terms in this expression determine alternative hypotheses. These hypotheses are ranked according to a given hypothesis evaluation criterion, LEF_h, and the *alt* best hypotheses are selected, where *alt* is a parameter provided to the program. The method is extended to inconsistent covering problem by introducing an event membership probability function. The selected hypotheses can be used as alternative generalizations of data, or arranged into an ensemble of classifiers to perform a form of boosting. The ALT method is general, and can thus be employed not only in concept learning, but also for generating alternative solutions to any general covering problem.

Keywords: Covering problem, Alternative hypotheses, AQ learning, machine learning, natural induction, data mining and knowledge discovery, knowledge mining, learning from examples.

Acknowledgments

The author thanks Ken Kaufman and Janusz Wojtusiak for helpful suggestions and comments on the earlier versions of this paper. This research was done in the Machine Learning and Inference Laboratory of George Mason University. Support for the Machine Learning and Inference Laboratory's research activities related to this research has been provided in part by the National Science Foundation under Grants No. DMI-9496192, IRI-9020266, IIS-9906858 and IIS-0097476; in part by the UMBC/LUCITE #32 grant. The findings and opinions expressed here are those of the authors, and do not necessarily reflect those of the above sponsoring organizations.

1 INTRODUCTION

From any non-trivial set of concept examples, it is usually possible to generate many alternative inductive generalizations of these examples, that is, inductive hypotheses. Such alternative hypotheses can be useful for a variety of practical applications of computational learning systems. For example, in medical decision making (diagnosis, drug prescription, or therapy assignment), some tests required by a given diagnostic procedure may be unavailable, and an alternative procedure would be necessary. Alternative hypotheses can also be used to increase the accuracy of classification decisions. This can be done through simple voting on decisions assigned by different hypotheses, or by weighted voting, as is typically done in boosting (e.g., Shapire and Singer, 1999).

The problem considered here is how to generate a set of alternative hypotheses that optimizes a given multi-criterion measure of hypothesis set quality. The hypotheses are assumed to be in the form of *attributional rulesets* (Michalski, 2004). The proposed method, A^q-ALT, or briefly, ALT, solves this problem by extending the classical A^q learning algorithm.

Before presenting the method in detail, let us start by outlining the main idea for readers that are familiar with the A^q algorithm. ALT proceeds in two steps. The first step proceeds as classical A^q, but from each star generated for some seed event, not one, but rather a set of complexes is selected, called an *elite for this seed.* An elite contains at most the *elitestar* complexes covering the seed that are evaluated as the best according to the $\text{LEF}_{\textit{star}}$ criterion, where *elitestar* is a user-provided parameter. Following the standard A^q algorithm (Michalski, 1969) and its version adapted to concept learning (Michalski, 1972), from each elite the best complex is selected, and events covered by it are removed from the set of positive concept examples (or target events) to be covered. This phase ends when all target events are covered by the selected complexes. Complexes in the set-theoretical union of elites are arranged into an *event covering table*, in which each positive target event is associated with the complexes that cover it. The concept of an elite is extended to apply to each event in the table, that is, an *elite of an event* is the set of complexes in the table covering this event.

The second phase of ALT determines all possible combinations of complexes whose union covers all events. This is done by logically multiplying out disjunctions of symbols denoting complexes in each elite, and determining an irredundant DNF expression. Each product in the so-obtained expression represents a complete and consistent hypothesis, or a cover of the target set against the contrast set. The generated hypotheses are ranked according to a given hypothesis evaluation criterion, LEF_{hyp}, and at most *alt* best are selected, where *alt* is a method parameter. Thus, the method seeks a collection of the “best” alternative hypotheses, not just any alternative hypotheses.

2 DESCRIPTION OF THE ALT METHOD

2.1 Notation and assumptions

Let $\mathcal{E}$ be an event space, and $\mathbf{E}^+$ and $\mathbf{E}^-$ be disjoint subsets of $\mathcal{E}$, called *target* and *contrast* datasets. Let $\mathcal{C}$ be a set of complexes, defined as predefined subsets of $\mathcal{E}$, such that for each event from $\mathcal{E}$, at least one complex covers it. In different problems, complexes have different meaning.

For example, in AQ programs for concept learning (programs based on the A^q algorithm), complexes are conjunctions of attributional conditions or selectors (Michalski, 2004). In minimization of Boolean expressions, complexes are conjunctions of literals (i.e., binary variables or their negation). In determining the minimum number of drugs needed to treat a given collection of diseases, complexes are individual drugs (assuming that each drug can treat more that one disease).

A general covering problem[*] is to determine a cover, COV($\mathbf{E}^+$| $\mathbf{E}^-$), of $\mathbf{E}^+$ against $\mathbf{E}^-$, which is a set of complexes whose set-theoretical union covers all events in $\mathbf{E}^+$**,** does not cover any event in $\mathbf{E}^-$, and optimizes a given criterion of cover quality, A quality criterion may be to minimize the number of complexes in the cover, or the total cost of the cover, when complexes are assigned different costs.

In the A^q algorithm for solving a general covering problem, the basic concept is that of a star, G(e **|** $\mathbf{E}^-$), of e against $\mathbf{E}^-$, defined the set of all possible maximally general complexes covering e and not covering any event in $\mathbf{E}^-$. While the ALT method places no restriction on the type of $\mathcal{E}$ and $\mathcal{C}$, we will assume here that $\mathcal{E}$ is spanned over multi-type attributes, and the target and contrast events are positive and negative examples of a concept whose general description is to be hypothesized. Without loss of generality, we will also assume that complexes are any subsets of the event space that are describable by a single attributional rule (Michalski, 2004), and a cover, COV($\mathbf{E}^+$| $\mathbf{E}^-$), is in the form of a set of such rules. Any cover that is a generalization of $\mathbf{E}^+$ also covers events that are not in $\mathbf{E}^+$, and is called a *hypothetical concept description*, or, briefly, a *hypothesis*.

Let LEF_{star} be a multi-criterion measure of quality of complexes in the star; and LEF_h be a multi-criterion measure of quality of a hypothesis. Finally, let *maxstar*, *maxelite*, and *alt* be user-provided control parameters that define the maximum number of complexes retained at any step of the star generation, the maximum size of the elite, and the maximum number of hypothesis to be generated by ALT, respectively.

The original A^q learning algorithm (the simplest version) for generating a cover, COV($\mathbf{E}^+$| $\mathbf{E}^-$), is presented in Figure 1 (based on [Michalski, 1969, 1971]).

```
Given $\mathbf{E}^+$, $\mathbf{E}^-$, $LEF_{star}$, maxstar, elitestar
1.  Select a seed event $e \in \mathbf{E}^+$.
2.  Generate a star G(e| $\mathbf{E}^-$).
3.  Select from the star the highest rank (best) complex, L.  according to $LEF_{star}$.
4.  Reduce $\mathbf{E}^+$ by removing from it examples covered by C.
5.  If $\mathbf{E}^+ = \varnothing$, stop; the collection of best complexes is a cover, COV($\mathbf{E}^+$| $\mathbf{E}^-$);
     otherwise, go to 1.
```

Figure 1: The original A^q algorithm (simple version).

The algorithm starts by randomly selecting a seed event, and then generates a star of it against the contrast (negative) events. The best (highest rank) complex according to the criterion LEF_{star} is

[*] We use the term "general covering problem" to distinguish it from a standard covering problem that does not have a contrast set.

selected from it, and events covered by it are removed from the set of target events, $\mathbf{E}^+$. A new seed event is then selected from the remaining target events, and the process repeats until the set of target events is empty, which means that the union of selected best complexes covers all original positive examples.

The ALT algorithm modifies the original algorithm in order to generate at most *alt* highest rank alternative covers according to the LEF_{hyp} criterion. It consists of two phases:

- Phase I determines the *primary* hypothesis and a collections of elites, and
- Phase II generates from a collection of elites at most *alt* highest rank hypotheses according to LEF_{hyp}, a predefined criterion of hypothesis quality.

Phase I is described in Figure 2. The basic idea of Phase I is to determine from each star generated not just one, the best complex, but the elite, that is, a set complexes of the highest rank according to the given criterion for evaluating complexes, LEF_{star}. Here are two examples of possible criteria for evaluating complexes:

$$\text{LEF}_{\text{star1}} = \ < \text{\#\_selectors, 0\%; selcost, 100\%}>$$
$$\text{LEF}_{\text{star2}} = \ < \text{\#\_target\_events\_covered, 25\%; \#\_selectors, selcost, 100\%}>$$

```
Given  $\mathbf{E}^+$, $\mathbf{E}^-$, $\text{LEF}_{\text{star}}$, alt, maxelite.

1. Select a seed event  $e \in \mathbf{E}^+$.
2. Generate a star $G(e|\mathbf{E}^-)$.
3. Select from $G(e|\mathbf{E}^-)$ the best complex, L, and the elite, EL, according to $\text{LEF}_{\text{star}}$.
     and store them in COV and EL-family, respectively.
4. Reduce $\mathbf{E}^+$ by removing from it examples covered by L.
5. If $\mathbf{E}^+ = \emptyset$, stop; COV is the premier hypothesis, and EL-family becomes
     an input to Phase II; otherwise, go to 1.
```

Figure 2: Phase 1 of the ALT algorithm.

The first criterion, $\text{LEF}_{\text{star1}}$, does not require counting the number of target events covered by a complex, i.e., #_target_events_covered. It selects the shortest complexes, thus the simplest and the most general. Such a criterion is particularly attractive in data mining where sets $\mathbf{E}^+$ and $\mathbf{E}^-$ can be very large. The second criterion, $\text{LEF}_{\text{star2}}$, is computationally more expensive, but may produce better solutions.

To control the complexity of the algorithm, the maximum size of the elite is limited by a predefined control parameter, *maxelite.* The algorithm proceeds and stops as the original one, ending with a cover $\text{COV}(\mathbf{E}^+|\mathbf{E}^-)$, and a collection of elites, an EL-family. Suppose that Phase I generated a family of elites, EL-family = $\{\text{EL}_1, \text{EL}_2, \ldots, \text{EL}_k\}$, where k is the number of stars generated. Phase II starts by creating an *event covering table*, in which columns correspond to events in $\mathbf{E}^+$, and rows correspond to complexes in the set-theoretic union, $\mathbf{U} = \cup\{\text{EL}_1, \text{EL}_2, \ldots, \text{EL}_k\}$, i= 1,2,…k, that is, to unique complexes selected from EL-family.

Each complex in $\mathbf{U}$ is matched against examples in $\mathbf{E}^+$ to determine examples covered by it. The table is then filled up according to the rule: If a complex covers an event, then the cell in the intersection of the column and the row corresponding to the event and the complex, respectively, is marked by 1; otherwise, it is marked by 0. Table 1 presents a very simple example of an event covering table (cells assigned “0” are left empty).

The event covering table is analogous to the prime implicant table used in the minimization of Boolean functions (e.g., McCluskey, 1956), and the proposed method for generating alternative hypotheses resembles the method for deriving irredundant expressions of Boolean functions described in (Petrick, 1956).

The main novelties here are that we are dealing here with complexes, which are more general concepts than prime implicants, and that **U** does not contain all the possible consistent and maximally general complexes that can be generated from the pair <$\mathbf{E}^+$, $\mathbf{E}^-$ > (training data), but only the highest rank complexes determined in Phase I. The last feature significantly simplifies the event covering table and the process of generating alternative hypotheses.

Complex	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8
C_1	1			1			1	
C_2		1	1		1		1	
C_3	1					1		
C_4		1	1				1	1
C_5		1			1			1

Table 1: A simple example of an event covering table.

The algorithm for Phase 2 is presented in Figure 3.

Given $\mathbf{E}^+$, EL-family, LEF_{hyp}, *alt.*

1. Create an event covering table, ECT, for the EL-family.
2. Determine events (columns) in ECT that are covered by a single complex. Remove columns corresponding to these events from ECT. Store complexes covering these events in COV_0.
3. For each event in the so reduced ECT, create a logical disjunction of symbols denoting complexes covering this event in the table.
4. Multiply out the created disjunctions to obtain an irredundant logic expression. The terms of this expression together with final value of COV_0 define alternative consistent and complete hypotheses.
5. Select from this expression at most *alt* best hypotheses according to LEF_{hyp}.

Figure 3: Algorithm for Phase II of the ALT method.

In Figure 3, LEF_{hyp} is a predefined multi-criterion formula for ranking hypotheses. The final value of COV_0 is called the *core cover*. More information on this formula is given in Section 3. The most difficult part of the algorithm for Phase II is Step 4. Let us discuss this step in detail.

Suppose that for an event e_i , i=1,2,3, … in ECT, a disjunction of complexes covering it is:

$$(C_{i1} \lor C_{i2} \lor \ldots \lor C_{ik}) \qquad (1)$$

To cover event e_i, one of the complexes in (1) must be present in any of the hypotheses. Therefore, the logical product

$$\bigwedge_{i} (C_{i1} \;\vee\; C_{i2} \;\vee\; \dots \;\vee\; C_{ik}) \qquad (2)$$

defines a set of alternative hypotheses. To determine these hypotheses, multiply out the product of disjunctions in (2). By applying absorption laws

$$C_1 (C_1 \;\vee\; C_2) \;=\; C_1 \quad \text{and} \quad C_1 \;\vee\; C_1 \; C_2 \;=\; C_1 \qquad (3)$$

in this process, an irredundant disjunction of conjunctions of symbols representing complexes is obtained:

$$\bigvee_{i} C_{i1} C_{i2} \; C_{i3} \; \dots \; C_{iz(i)} \qquad (4)$$

Each set of complexes included in a single product, $C_{i1} C_{i2} \; C_{i3} \; \dots \; C_{iz(i)}$, i =1, 2, 3, …, together with set COV_0, constitutes an alternative hypothesis.

The final step is to select from the obtained collection of hypotheses a set of best hypotheses according to the multi-criterion LEF_{hyp}. An example of such a criterion specification is presented in the next section.

3 AN ILLUSTRATION

To illustrate the ALT method, consider the event covering table (ECT) in Table 1. The core cover is:

$$COV_0 = \; \{C_1 , C_3 \}$$

because e_4 is covered only by C_1, and e_6 is covered only by C_3. By removing from ECT these complexes and the events covered by them, the following reduced event covering table is obtained (Table 2).

Complex	e_2	e_3	e_5	e_7	e_8
C_2		1	1	1	
C_4	1	1		1	1
C_5	1		1		1

Table 2: Reduced event covering table.

From that table, the product is generated:

$$(C_4 \;\vee\; C_5)(C_2 \vee C_4)(C_2 \vee C_5)(C_2 \vee C_4)(C_4 \vee C_5) = \; C_4 \; C_2 \;\vee\; C_4 \; C_5 \;\vee\; C_2 \; C_5 \qquad (5)$$

After multiplying out (5), applying the absorption laws, and including COV_0 in the resulting expression, the following irredundant DNF expression is obtained:

$$C_1 C_3 C_4 C_2 \;\vee\; C_1 C_3 \; C_4 C_5 \;\vee\; C_1 C_3 \; C_2 C_5 \qquad (6)$$

Each product in (6) corresponds to one cover or hypothesis. Presenting these covers as sets of complexes, we obtain the following collection of alternative hypotheses:

$$\{C_1 , C_3 , C_4 \; , C_2 \}, \{ C_1 , C_3 \; , \; C_4 \; ,C_5 \}, \; \{ C_1 , C_3 , \; C_2 , C_5 \} \qquad (7)$$

These hypotheses are evaluated according to the LEF_h multi-criterion formula, and ordered from the best to the worst. Suppose that *alt* = 2, and

$$LEF_{hyp} = \text{<No\_of\_rules, 10\%; No\_of\_conditions, 100\%>} \tag{8}$$

where No_of_rules is the total number of rules in a hypothesis, and No_of_conditions is the total number of conditions in the hypothesis. The criterion (8) first ranks hypotheses according to the number of rules, and selects a subset in which the longest hypothesis is not more than 10% longer than the shortest one. Next, it ranks the rules in the selected set according to the total number of conditions occurring in the rules of each hypothesis, and keeps them all (because the tolerance is 100%). One could also evaluate the cost of evaluating the conditions in rules, but for simplicity we will ignore this factor here.

To evaluate these hypotheses using the above LEF_{hyp}, suppose that complexes C_1, C_2, C_3, C_4, C_5 have 2, 5, 7, 1, and 8 conditions, respectively. Table 3 characterizes the obtained hypotheses in terms of No_of_rules and No_of_conditions.

Number	Hypothesis	No_of_rules	No_of_conditions
1	$\{C_1, C_3, C_4, C_2\}$	4	16
2	$\{C_1, C_3, C_4, C_5\}$	4	18
3	$\{C_1, C_3, C_2, C_5\}$	4	22

Table 3: Complexity of complexes in alternative hypotheses.

Because all hypotheses have the same number of rules, their ranking is decided by the number of conditions in them. At most *alt* = 2 best hypotheses can be chosen, thus hypotheses 1 and 2 are selected as the output set of alternative hypotheses.

4 DEALING WITH AN INCONSISTENT COVERING PROBLEM

The method presented a solution of a *consistent* covering problem, in which the target set $\mathbf{E}^+$ and the contrast set $\mathbf{E}^-$ are disjoint. This section extends the method to an *inconsistent covering problem*, in which sets $\mathbf{E}^+$ and $\mathbf{E}^-$ have a non-empty intersection. We assume that each event in the intersection can be assigned a probability of belonging to $\mathbf{E}^+$ (its complement is assumed to be the probability of belonging to $\mathbf{E}^-$). Such a probability can be estimated by the ratio of the number of times the event is assigned to $\mathbf{E}^+$ to the total number of occurrences of that event. The presented method is based on ideas introduced in (Michalski and McCormick, 1971).

Let f be a event membership probability E → [0, 1, ?], such that

$$\mathbf{E}^+ = \{e \in \mathcal{E} : f(e) = 1\}$$

$$\mathbf{E}^- = \{e \in \mathcal{E} : f(e) = 0\}$$

$$\mathbf{E}^\phi = \{e \in \mathcal{E} : 0 < f(e) < 1\}$$

$$\mathbf{E}^? = \{e \in \mathcal{E} : f(e) = ?\} = \mathcal{E} \setminus (\mathbf{E}^+ \cup \mathbf{E}^- \cup \mathbf{E}^\phi) \tag{9}$$

where "?" means that the value of f(e) is unknown for event e. Events in $\mathbf{E}^?$ are not used for learning Events in $\mathbf{E}^\phi$ are inconsistent, because there is a non-zero probability that they belong to the target set, and a non-zero probability that they belong to the contrast set.

Let us arrange the inconsistent events in descending order of f(e), that is, from those most likely belonging to the target set to the most likely belonging to the contrast set. Assuming a threshold $\lambda \in [0, 1]$, we define:

$$\mathbf{E}^{+\lambda} = \{e \in \mathbf{E}^{\phi} \mid f(e) \geq \lambda\}$$
$$\mathbf{E}^{-\lambda} = \{e \in \mathbf{E}^{\phi} \mid f(e) < \lambda\} \tag{10}$$

If all events are arranged along a horizontal axis in descending order of values of f(e), then the event membership function, f(e), may look like in Figure 4.

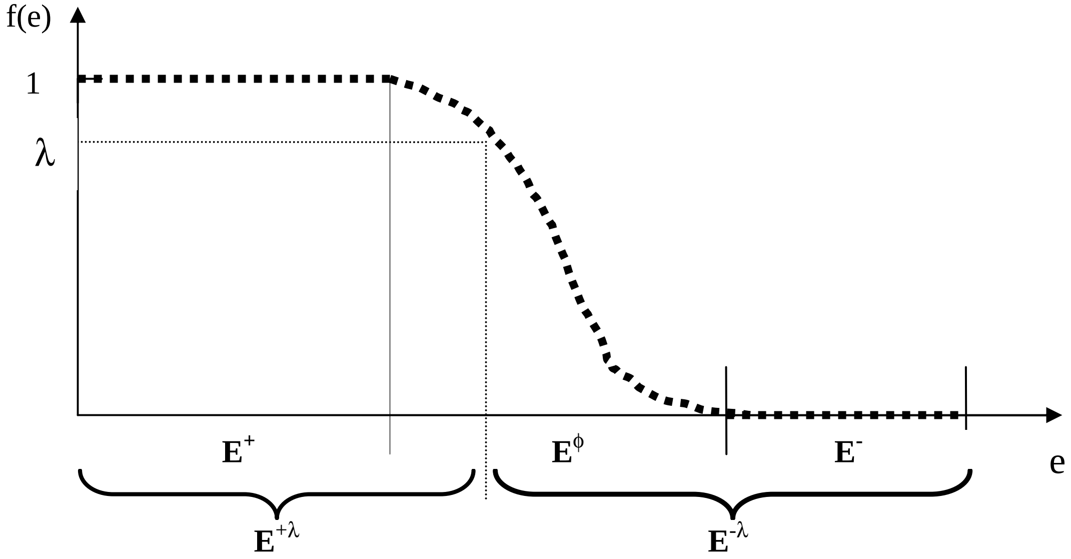

Figure 4: The event membership probability function (the probability of e belonging to $\mathbf{E}^{+}$)

For any specific value of λ, the sets $\mathbf{E}^{+\lambda}$ and $\mathbf{E}^{-\lambda}$ are disjoint (Figure 4), so the problem reduces to the consistent covering case, and the previously described algorithm can be applied. If $\lambda = 0$, all inconsistent events are treated as belonging to $\mathbf{E}^{+}$**.** If $\lambda = 1$, all inconsistent events are treated as belonging to $\mathbf{E}^{-}$**.** If $\lambda = 0.5$, then events with probability f(e) greater than or equal to 0.5 are assumed to belong to the target class. One can also ignore events in $\mathbf{E}^{\phi}$, and assume that $\mathbf{E}^{+\lambda} = \mathbf{E}^{+}$ and $\mathbf{E}^{-\lambda} = \mathbf{E}^{-}$.

All four of these possibilities have been implemented in the AQ21 learning program (Wojtusiak, 2004). To make the choices easy to remember, instead of λ a user-defined parameter “ambiguity” is used, which can be set to one four values, each corresponding to one of the above choices. Table 4 presents these choices and corresponding values of λ.

The last row describes a method which determines a pair of covers, one for $\lambda = 0$ and one $\lambda = 1$. The two covers correspond to what in rough set theory are called the upper bound and the lower bound approximations of the concept represented by the target and contrast datasets, respectively (Pawlak, 1991).

λ	*Ambiguity*	**Explanation**
0	IncludeInPos	Ambiguous events are included in the target set (are treated as positive examples for concept learning) and removed from the contrast set.
1	IncludeInNeg	Ambiguous events are included in the contrast set (are treated as negative examples) and removed from the target set.
–	IgnoreForLearning	Ambiguous events are removed (ignored) from the data.
0.5	IncludeWhere MostFrequent	Ambiguous events are included in the set in which they are most frequent and removed from the other (target or contrast) set.
0 & 1	CreateTwoCovers	Two hypotheses are generated, one with ambiguous events included in the target set, and the other with them included in the contrast set.

Table 4: Methods for handling inconsistent covering problems implemented in AQ21.

5 DETERMINING BEST ALTERNATIVE CLASSIFIERS

Choosing hypotheses only on the basis of their complexity may not be sufficient. This is so because such a criterion does not take into consideration the degree of similarity or other relationships among hypotheses. Clearly, when selecting alternative hypotheses, it is desirable to select not only those that are the simplest but also those that are most different from each other, or have some other desirable properties.

Let us consider first the case of two alternative hypotheses (rulesets). One measure of the difference between two hypotheses is an *attribute disjointness* or, briefly, *a-disjointness*, which is the total number of attributes that are not shared by these hypotheses. Another criterion, called *selector disjointness*, or, *s-disjointness*, could the total number of selectors in two hypotheses that associate a different reference with the same attribute. A measure of s-disjointness may also be extended to distinguish different types of relations between the references in the selectors, such as complete disjointness, partial disjointness, and subsumption.

To evaluate a collection of alternative rulesets (hypotheses), one can measure the average of a- and s-disjointness between all pairs of hypotheses in the collection. Thus, we can introduce an additional criterion that ranks different collections of hypotheses, LEF_{coll} defined as:

$$LEF_{coll} = \langle \text{a-disjointness}, \tau;\ \text{s-disjointness} \rangle \qquad (11)$$

where τ is a tolerance on the a-disjointness measure. The criterion, LEF_{coll}, could be combined with a criterion for evaluating individual hypotheses, LEF_{hyp}, such as (8).

In the multi-class case, a set of alternative rulesets is learned for each class. A collection of alternative rulesets for all classes is called an *alternative ruleset assembly* or *ARA*. By selecting one hypothesis from among alternative hypotheses in the ARA for each class, one can generate alternative classifiers (families of rulesets).

ARAs thus allow one to generate a set of different classifiers for a given problem. If there are many classes, and several hypotheses for each decision class (concept), the set of all alternative ruleset families (classifiers) can be very large. One can then use an additional criterion, LEF_{clsf}, for ranking classifiers, and select a subset of the best ones.

The LEF_{clsf} criterion for ranking classifiers can involve such measures as the total number of rules in the classifier (#rules), the total number of selectors (#sel), a classifier complexity measure (as defined in the ITG program employed in AQ21; Wojtusiak, 2004), and other such measures. One can also select a subset of the testing set, determine the predictive accuracy of the classifiers on that subset, and then select the best performing classifiers. Once alternative classifiers are ranked according to LEF_{clsf}, the best of them can be selected for actual use in decision making.

6 BUILDING A CLASSIFIER ENSEMBLE FROM ALTERNATIVE HYPOTHESES

A collection of different hypotheses can be used for improving predictive accuracy in classification. The simplest method is to apply alternative hypotheses to a given event, and then make a decision according to the majority voting rule. In order to avoid the possibility of a tie, an odd number of alternative hypotheses should be generated for the target dataset.

A more advanced method would be weighed voting, in fashion resembling boosting (Schapire and Singer, 1999; Hastie, Tibshirani and Friedman, 2001). In this method, different hypotheses for each class are assigned different weights. A weighted sum of the degrees to which hypotheses are satisfied by an event would generate the score given for the decision associated with hypotheses of a given class. The decision with the highest score would be the output of the procedure.

The weight given to a hypothesis could be assigned in different ways. If the hypotheses are approximate with regard to the training data (partially inconsistent and/or incomplete), the weight of the hypothesis could be based on the accuracy of the hypothesis on the training set.

An alternative method, applicable also in the case of consistent and complete hypotheses, would be to split the original training set into a primary and secondary set. The hypotheses would be learned from the primary set, and then their accuracy would be estimated on the secondary training set. The so-estimated accuracy of the hypotheses would be used to assign the weights to different hypotheses for the purpose of weighted voting in classifying events in the testing set.

7 DETERMINING MULTI-CLASS COVERS

The above sections described an algorithm for generating alternative hypotheses for a target dataset in the context of a contrast dataset, that is, for solving a two-class covering problem. The method can be generalized to a multi-class covering problem in a straightforward fashion.

When there are many classes, the algorithm is repeated for each class, taking each class as the target set, and the union of the remaining classes as the contrast set. An alternative method to this *parallel* multi-class cover, is to generate a *sequential* multi-class cover.

To generate a sequential multi-class cover, classes are ordered into a sequence according to some criterion, for example, from the largest to the smallest. Suppose C_1, C_2, ..., C_m is such an ordered sequence. First, a cover is generated taking the first class as the target set and remaining classes as the contrast set, that is, COV(C_1 | C_2,...., C_m). Subsequently, covers COV(C_2 | C_3,...., C_m), ..., COV(C_{m-1} | $C_{m)}$ are generated. Thus, there is no cover for the last class.

To classify an event, it is first matched against the first cover, then, if it does not match it, it is matched against the second cover, and so on. If it does not match any cover, up to that of the *m*-1st class, it is classified to the *m*th class.

Because individual descriptions in a sequential multi-class cover are created using consecutively smaller contrast sets, such a cover is always simpler than a parallel multi-class cover. A price for this advantage is that classifying events using such a cover requires matching covers of individual classes sequentially until one is matched, thus it cannot be done in parallel. In addition, an interpretation of individual cover descriptions is more difficult because each class description is a conjunction of the description of the given cover and the negation of the descriptions of predecessor covers.

8 CONCLUSION

The ALT method was developed with the primary purpose of creating alternative hypotheses in concept learning using the A^q algorithm. The method is, however, general, and can be applied for determining alternative solutions to any covering problem. It can thus be used for determining alternative hypotheses using other rule learning methods developed in machine learning or data mining, as well as for solving covering problems in other application domains, such as the optimization of communications networks, the minimization of switching systems, determining a minimal set of drugs needed for treating a given set of diseases, or designing optimal psychometric testing (Hammer and Rader, 2001).

The method concerns the case of generating alternative hypotheses that are consistent and complete with regard to the input data (training set). It can, however, be easily extended to the case of generating alternative approximate hypotheses. In this case, the AQ learning program should be run in *pattern discovery* mode, rather then *theory formation* mode, as assumed in this paper, and the elite set should consist of best approximate complexes that may cover some events in the contrast set $\mathbf{E}^-$, rather then being fully consistent.

The paper also described several methods for solving inconsistent covering problems in which the target set and the contrast set overlap. Such a situation occurs in practical problems when the set of attributes for describing events is insufficient. The presented method has been implemented in the AQ21 learning environment. Results from an experimental investigation of the method will be published in a separate report.

REFERENCES

Hammer, P. and Rader, D.J., “Maximally Disjoint Solutions of the Set Covering Problem,” *Journal of Heuristics*, 7:131-144, Kluwer Academic Publishers, 2001.

Hastie, T., Tibshirani, R. and Friedman, J., The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2001.

McCluskey E. J., “Minimization of Boolean Functions,” Bell System Technical Journal, vol. 25, pp.1477-1444, November 1955.

Michalski R.S., “On the Quasi-Minimal Solution of the General Covering Problem,” Proceedings of the V International Symposium on Information Processing (FCIP 69) (Switching Circuits) , Vol. A3 , Bled, Yugoslavia, pp. 125-128, October 8-11, 1969.

Michalski R.S., “A Geometrical Model for the Synthesis of Interval Covers,” Report No. 461, Department of Computer Science, University of Illinois, Urbana, June 24, 1971.

Michalski R.S., “A Variable-Valued Logic System as Applied to Picture Description and Recognition," in F. Nake and A. Rosenfeld (eds.), Graphic Languages, North-Holland Publishing Co., 1972.

Michalski R.S., “ATTRIBUTIONAL CALCULUS: A Logic and Representation Language for Natural Induction,” Reports of the Machine Learning and Inference Laboratory, MLI 04-2, George Mason University, Fairfax, VA, April, 2004.

Michalski R.S. and McCormick, B. H, “Interval Generalization of Switching Theory,” Report No. 442, Department of Computer Science, University of Illinois, Urbana, May 3, 1971.

Pawlak, Z, Rough Sets: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, 1991.

Petrick S. R., “A Direct Determination of the Irredundant Forms of a Boolean Function from the Set of Prime Implicants,” AFCRCTR-56-110, Air Force Cambridge, Research Center, Cambridge, Mass., April, 1956.

Schapire, R. and Singer, Y., “Improved Boosting Algorithms Using Confidence-rated Predictions,” Machine Learning, 37, Kluwer Academic Publishers, pp.297-336, 1999.

Wojtusiak, J., “AQ21 User’s Guide,” Reports of the Machine Learning and Inference Laboratory, MLI 04-3, George Mason University, Fairfax, VA, September, 2004.

A publication of the *Machine Learning and Inference Laboratory*
School of Computational Sciences
George Mason University
Fairfax, VA 22030-4444 U.S.A.
http://www.mli.gmu.edu

Editor: R. S. Michalski
Assistant Editor: K. A. Kaufman

The *Machine Learning and Inference (MLI) Laboratory Reports* are an official publication of the Machine Learning and Inference Laboratory, which has been published continuously since 1971 by R.S. Michalski's research group (until 1987, while the group was at the University of Illinois, they were called ISG (Intelligent Systems Group) Reports, or were part of the Department of Computer Science Reports).