The PROMISE Method For Selecting Most Relevant Attributes
For Inductive Learning Systems

by

Paul W. Baim

September 1982
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

ISG 82-1
File No. UIUCDCS-P-82-898

This work was supported, in part, by a National Science Foundation
Grant No. MCS-82-05166 and, in part, by an Office of Naval Research
Grant No. N00014-82-K-0186.
Acknowledgements

I would like to thank the following people for their invaluable help: Professor R.S. Michalski for his constant encouragement and critical eye; Bob Stepp for his help in editing and proofreading this paper.
### Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>2. Methodology</td>
<td>2</td>
</tr>
<tr>
<td>2.1 The PROMISE Algorithm</td>
<td>3</td>
</tr>
<tr>
<td>3. Experimentation</td>
<td>4</td>
</tr>
<tr>
<td>3.1 Experiment I</td>
<td>4</td>
</tr>
<tr>
<td>3.2 Experiment II</td>
<td>7</td>
</tr>
<tr>
<td>3.3 Experiment III</td>
<td>8</td>
</tr>
<tr>
<td>4. Conclusions</td>
<td>8</td>
</tr>
<tr>
<td>5. References</td>
<td>10</td>
</tr>
</tbody>
</table>
1. Introduction

An important problem in the development of inductive learning systems is the construction of decision rules based only on examples of one or more classes of objects in the form of vectors of values of a set of attributes of the objects (e.g., chocolates may be divided into nut-centers, creams, fruit-centers, etc. and one of the attributes may be: "Is the center fruit-flavored?"). Such systems are needed in areas such as agriculture, medicine, and mineral prospecting where they function as "expert" diagnosticians or mineralologists. These expert systems help to alleviate the problems caused by the lack of a sufficient number of human experts to satisfy the need for them.

In logic-based inductive learning systems, a data analyst assembles a collection of events (from one or more classes) each described by a vector of values of attributes. The values of each of the attributes will, to some extent, distinguish one or more classes from the others (in the chocolates example, the fruit-flavor attribute distinguishes creams and fruit-centers from the other classes but not from each other and this attribute alone does not distinguish all creams). Some of the attributes may be less relevant than others because their values do not distinguish any of the classes from the others. For optimum performance of the inductive system and to minimize the intricacy of the generated rules which distinguish the classes of events, the smallest possible set of attributes should be used.

The problem of selecting a minimum set of attributes to describe objects for the purpose of inductive learning is traditionally approached using various methods of attribute selection such as factor analysis, multidimensional scaling, and linear transformation which were developed in the fields of pattern recognition and statistical decision theory. These methods involve statistical or information-theoretic measures which identify the principal attributes in the data-set [1,2]. These methods are effective when the values of the attributes are partially ordered; however, purely mathematical analysis of nominal attributes is often misleading. In fact, using the methods mentioned above, "it is almost impossible to formulate general guidelines regarding the selection of physical [attributes (such as color or fragrance)] and structural [attributes]" [3,p245] (such as shape or texture). To analyze nominal attributes mathematically, numeric values must be
assigned to each possible nominal value of the attribute (e.g., color might be encoded as: red=1, yellow=2, blue=3, etc.). When statistical methods are applied to such attributes, the results from the statistical methods may vary if the ordering of the values changes (e.g., red=2 and yellow=1). Although the actual values of the attribute have not changed, the measures derived from them have. This is clearly erroneous.

The difficulty of performing satisfactory attribute selection using statistically-based methods may be overcome by analyzing the observed values of the attributes using a logical method rather than a statistical one. Such a system could deal effectively with nominal and ordinal attributes. The need for such an attribute-selection method is revealed by considering the performance of the inductive-learning program AQ11 [4]. The computation time necessary to process a set of events is large when the data-set contains many attributes. A computationally inexpensive means to rapidly eliminate unpromising attributes would reduce the computation time required. The method described in this paper accomplishes all of these goals.

Three experiments were performed to explore the performance of this method when applied to various data-sets. When a significant number of unpromising attributes are included in the data-set, as would be the case when many attributes of unknown relevancy are used, the processing time observed for AQ11 is significantly reduced when the data is pre-processed using the method described here to eliminate the irrelevant attributes.

2. Methodology

A data-set composed of vectors of attribute values may be described as a table where columns contain the values of a particular attribute in each of the vectors and rows contain the values of each of the attributes in a specific vector. Events (vectors of attribute values) can have special relationships with other events in the data-set. When the values of the attributes of one event are the same as those of one or more events in one or more other classes, this set of identical events will be termed a Cross-Class Equivalency (CCE). When such identical events occur within a single class, the set of identical events is termed an In-Class Equivalency (ICE).

A table may be reduced in size by the projection operation. A projection onto a subset of
attributes is a table containing all of the original events but only the specified subset of the original set of attributes. A projection may contain CCEs and ICEs when the original table had none.

The algorithm described below is based on three assumptions about data-sets of interest in inductive-learning systems. First, the relevance of an attribute (termed the promise of the attribute) is assumed to vary inversely with the number and size of CCEs in the projection of the data-set on just that attribute (although PROMISE may be used to evaluate more than one attribute at a time). Second, the occurrence of two or more attributes which are unpromising individually but are promising when considered together is rare (i.e., the promise of an attribute may be estimated without considering the interrelationships between attributes). Third, in a large data-set with many attributes, a significant portion of the attributes may be irrelevant.

2.1 The PROMISE Algorithm

The promise-evaluation algorithm PROMISE works as follows:

Algorithm PROMISE: This algorithm determines the promise of an attribute by computing a "cost" based on the loss of class-distinguishing information due to CCEs within the projection of the data-set onto the current attribute of interest. It does this by selectively removing events from CCEs and re-evaluating the data-set until no CCEs remain. The algorithm returns a value, p, which is a measure of the promise of the attribute where p ranges from 0 for a perfect attribute (without CCEs) to m-1 for the worst case (when the attribute has the same value in all events) where m is the number of distinct classes in the data-set. The initial value of p is 0.

1. Project the data-set on the attribute.
2. Locate an event, e, that is a member of a CCE.
3. Find the largest class, c, with the smallest ICE within the CCE which contains the event found in step 2.
4. Increase p by the size of the ICE in the class found in step 3 divided by the size of the class.
5. Remove the events in the ICE accounted for in step 4 from class c.
An example will demonstrate the functioning of the algorithm. If we have the data-set of 3 classes and 2 attributes given in Table 1, we would compute the values of $p$ for attributes $x_1$ and $x_2$ as follows: Attribute 1 has a CCE for value 1 in classes 1 and 2. Since the number of 1-values for attribute 1 is least in class 2, the value of $p$ increases from 0 (the initial value) to 0.5 (as per steps 3 and 4 above). After the events in the ICE in class 2 are removed, no CCEs remain. Attribute 2 has a CCE for value 1 in all 3 classes. Since the number of 1-values is least in both classes 2 and 3, and class 2 is larger than class 3, we process the 1s in class 2 first. This increases $p$ from 0 to 0.5. Once the 1s in class 2 have been dealt with, they are removed. Therefore, the next CCE contains the 1s in classes 1 and 3. Since the least 1s are in class 3, we increase $p$ from 0.5 to 1.5. After removing the 1s from class 3, no CCEs remain. Looking at the final promise values for attributes 1 and 2 (0.5 and 1.5 respectively), we see that attribute 1 is better than attribute 2 as we might expect from examining the data-set.

3. Experimentation

Three experiments were performed to explore the behavior of the PROMISE algorithm and to determine its effectiveness when used to pre-process data to be further processed by AQ11 (AQ11 and PROMISE are implemented in Pascal on a Cyber 175).

3.1 Experiment 1

The PROMISE algorithm can be used to analyze more than one attribute at a time by considering the values of several attributes in an event as a single compound attribute. To test the second assumption, the "Animals" data-set, described in [5], was processed by PROMISE. First, the projections of the data-set onto single attributes were evaluated, and the attributes were arranged in order of increasing value of $p$. Next, all projections on pairs of attributes were evaluated and the results were analyzed as follows: The combinations were ranked by increasing value of $p$ and the range of observed values of $p$ was divided into 8 equal-sized sub-ranges. The number of occurrences of a particular attribute in the set of pairs with promise values in a subrange is expressed as a percentage of the total number of occurrences of all attributes in the subrange and plotted as in Figure 1. A promising attribute should exhibit greater participation in
Fig 1. Graph of the % participation of an attribute across several ranges of p-value (see text).

<table>
<thead>
<tr>
<th></th>
<th>x_1</th>
<th>x_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>c_1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

| c_2 | 1   | 1   |
|     | 2   | 2   |

| c_3 | 3   | 1   |

Table 1. A sample data set used to illustrate the functioning of PROMISE.
Table 1. Attributes ranked by promise values derived from analysis of combinations of various numbers of attributes.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>best</th>
<th>worst</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Attributes</td>
<td>$x_1 \ x_9 \ x_6 \ x_{10} \ x_3 \ x_2 \ x_4 \ x_8 \ x_{13} \ x_5 \ x_{11} \ x_{12} \ x_7$</td>
<td></td>
</tr>
<tr>
<td>Pairs</td>
<td>$x_1 \ x_9 \ x_6 \ x_{10} \ x_4 \ x_{13} \ x_3 \ x_2 \ x_8 \ x_{11} \ x_5 \ x_{12} \ x_7$</td>
<td></td>
</tr>
<tr>
<td>Triples</td>
<td>$x_1 \ x_9 \ x_6 \ x_{10} \ x_2 \ x_3 \ x_4 \ x_8 \ x_{13} \ x_{11} \ x_5 \ x_{12} \ x_7$</td>
<td></td>
</tr>
<tr>
<td>Quadruples</td>
<td>$x_1 \ x_9 \ x_6 \ x_{10} \ x_4 \ x_{13} \ x_2 \ x_3 \ x_8 \ x_{11} \ x_5 \ x_7 \ x_{12}$</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Attributes ranked by promise values derived from analysis of combinations of various numbers of attributes.

<table>
<thead>
<tr>
<th>best</th>
<th>worst</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_6 \ x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_8 \ x_9 \ x_7$</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Attributes from a test data set ordered by individual promise values.
the low-valued sub-ranges and unpromising attributes should exhibit greater participation in high-valued sub-ranges. When linear regression analysis is applied to these plots, promising attribute "profiles" should have more negative slopes and greater y-intercepts and unpromising attribute profiles should have more positive slopes and lower y-intercepts. Such an analysis was performed for all combinations of three and four attributes as well. The resulting attribute rankings are shown in Table 2. The table shows that the promise value measured for each attribute is independent of the interactions between attributes in this data-set because the rankings are fundamentally the same independent of the size of the groupings used. Therefore, the values indicated by PROMISE for individual attributes are independent of the interactions between attributes in this experiment, and can be used to order the attributes.

3.2 Experiment II

The data-set used for experiment II was designed so that attributes with widely varying degrees of promise were included. The values of one attribute were arranged so that the classes could be distinguished by the value of that attribute alone. Another attribute had the same value in all events. Two more attributes were pseudo-random, and the rest differentiated the classes to varying degrees. The data comprised five classes (3 with 8 events, 2 with 5 events). The result of the analysis of the attributes individually is shown in Table 3. The ordering of the attributes by promise value (Table 3) matches the order of attribute relevancy designed into the data-set. The results show that PROMISE evaluates attributes based on their ability to distinguish classes in the data-set. This indicates that a set of attributes which uniquely characterizes each class in a data-set and contains few extraneous variables can be obtained by examining attributes beginning with the most promising attributes and adding more attributes in order of decreasing promise until the projection of the data-set on the attributes contains no CCEs. For example, when the most promising attribute is excluded, the next three attributes can distinguish the classes uniquely (two of them comprise the minimum number that do so in this data-set when the best attribute is excluded).
3.3 Experiment III

Experiment three was undertaken to determine the magnitude of the computational advantage, if any, realized by removing as many extraneous attributes as possible using PROMISE before the data is processed by AQ11.

The data for the third experiment was the same data used in experiment II with the most-promising attribute removed and a number of pseudo-random attributes added (42 in this experiment) to simulate the common occurrence of data-sets which include no perfect attributes and many irrelevant ones. The attributes were processed by PROMISE and ordered by increasing promise value. Because the three most promising attributes are not distinguished by different p values, they were all accepted initially and the projection of the data onto those three attributes was evaluated by PROMISE and found to have no CCEs (p=0). Next, the projected data-set was processed by the inductive-learning program AQ11 which derived rules to discriminate the classes. The rules derived from the projected data-set were identical to those derived when the entire data-set was processed by AQ11. Table 4 shows a comparison of the time needed to derive the same set of rules using AQ11 on the entire data-set and using AQ11 after PROMISE determined the 3 best attributes. In this instance, a computation-time reduction of 88% was observed.

4. Conclusions

A method to select relevant attributes for inductive learning has been developed. The method has been shown to be useful in reducing the time necessary to process a raw data-set in order to derive decision rules which differentiate given classes of events within the data-set. Experiments have shown that the method estimates the promise of an attribute without need for consideration of the attribute's interactions with other attributes. In addition, significant processing time reduction in an inductive-learning program has been demonstrated when the method is used to pre-process a data-set containing many irrelevant attributes. It overcomes the main limitations of statistically-based attribute selection methods when they are applied to nominal attributes, while exhibiting accurate discrimination of promising and unpromising
attributes when dealing with the types of data-sets encountered when applying inductive-learning techniques.

Evaluating groups of attributes using PROMISE may reveal strong interrelationships between attributes which could serve as a basis for combining the attributes using various logical and/or mathematical operators to produce new, derived attributes. The production of such derived attributes from simple ones has been termed constructive induction [6]. For example, the length, height, and width of an object may be important attributes for distinguishing the class of object. This may imply that the product of these variables (the volume) is a powerful class-distinguishing attribute. Volume is a constructively-induced attribute in this case. Once a new attribute has been constructed, its promise can be evaluated using PROMISE.

5. References

The PROMISE Method For Selecting Most Relevant Attributes For Inductive Learning Systems

Paul W. Baim

Department of Computer Science
University of Illinois
Urbana, IL

National Science Foundation
Office of Naval Research
Washington, DC
Arlington, VA

One of the failings of conventional attribute-selection methods is their inability to properly evaluate and select most relevant nominal and ordinal attributes. This paper describes an algorithm called PROMISE that can evaluate a large number of attributes and select a small subset which are sufficient to fully characterize the data-set. Three experiments are described which have demonstrated the performance of this algorithm when used to pre-process a set of attributes for use with an inductive-learning program.

Key terms: Inductive Learning, Attribute Selection, Data Analysis, Constructive Induction