CONTROL
AND CYBERNETICS
VOL. 18 (1989) No 3-4

INDUCTIVE LEARNING OF THE MINIMUM LENGTH CLASSIFICATION RULES WITHOUT RISKING COMBINATORIAL EXPLOSION

by

C. IWAŃSKI
G. SZKATUŁA

Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw, Poland

Inductive learning algorithms are threatened by combinatorial explosion.To avoid this difficulty the algorithms are either recommended for the problem with rather small number of attributes, Quinlan (1983),or a number of examples is limited, Garis (1988), or else some arbitrary constraints are introduced during the learning process, Michalski (1983). In this paper an inductive learning algorithm is described, which produces the minimum length classification rules (in the sense of numbers of attribute-value pairs). The algorithm works in polynomial time and the number of attributes and/or examples need not to be limited. Some results obtained for medical data are presented.

KEY WORDS: Machine learning, Learning by example, Combinatorial explosion, Integer programming.

Introduction

Inductive learning algorithms are threatened by combinatorial explosion. Consider, for example a single step of Michalski's (1983) algorithm. Let n be the number of negative examples, k the number of attributes and s the average number of selectors (attribute-value pairs) in the discriminant set. In such a case about s^n complexes can be generated, where $1<s\leq k$. Usually, k is in the tens and n is in the hundreds. Thus to avoid combinatorial explosion some arbitrary constraints must be introduced e.g. on the number of complexes, but such solution can be far from the most preferable. Usually the preference criterion is the length of classification rule (number of selectors in disjunction of consistent complexes). Quinlan's (1983) algorithm is recommended for the problem with rather small number of attributes. Similar situation occurs with other algorithms. Hugo de Garis's (1988) modification of Michalski's algorithm also includes a remark that in situation risking combinatorial explosion the number of positive examples for the class under discussion should be randomly reduced.

In this paper a modified Michalski's algorithm is proposed, which avoids combinatorial explosion without introducing any constraints during performance of algorithm or decreasing the number of data. The main idea is to replace the most time consuming part of forming consistence complexes to select the most preferable one by solving the simple problem of integer programming. Generally it is an NP-hard problem.However,it is a well known problem and there are many efficient heuristic algorithms giving useful solutions and working in polynomial time (in our case with respect to the number of examples and attributes). Besides that a special,not random selection of a positive example for each iteration is introduced. This considerably accelerates and improves the inductive process. The remainder of the paper is organized as follows. In Section 2 Michalski's approach called star methodology is described.

Section 3 introduces the modification to original algorithm. In Section 4 the method is illustrated by means of medical data. Finally, in Section 5 some conclusions are included.

2. Michalski's star approach

Inductive learning is the process of inferring classification rules from examples. There are many approaches in this area of machine learning. The most important ones are the following methods based on: searching version space Mitchell (1982), star methodology Michalski (1983) and the principle of maximizing expected information Quinlan (1983). Most often the star methodology approach is applied e.g. Michalski and Chilausky (1980), Michalski and Stepp (1983), Shaw (1987). An excellent review can be found in Ditterich et al. (1981).

In inductive learning process an example can be seen as a collection of attribute-value pairs and can be written in the variable-valued logic mode proposed by Michalski (1973) e.g.

$$\text{example} = [A_1\ r_1\ v_1]\ [A_2\ r_2\ v_2]\ \dots\ [A_k\ r_k\ v_k],$$

where A_i is an attribute name, r_i is a relation and v_i is a value from value set of A_i for $i=1,2,\dots,k$.

A form $[A_i\ r_i\ v_i]$ is called rational statement or selector. For example

[sex = man]   [size = medium]

[color = brown]   [weight = heavy]

are selectors with values represented by linguistic terms. A typical classification rule might take the form of

$$[A_2 r_2 v_2]\ [A_4 r_4 v_4] \cup [A_7 r_7 v_7] \Rightarrow [\text{Class} = \text{class}_3],$$

where premise part of classification rule is a disjunction of conjunctions of selectors.

The inductive learning process for multiple concepts works under assumption that there exist teachers/experts who can correctly classify training examples in classes. The training examples for which the concept description is to be found are called positive. The remaining ones are called negative. The goal of the algorithm is then to find a classification rule that correctly describes all positive examples and does not describe any negative examples. The basic role in Michalski's approach is played by definition of a consistent complex. A complex is a conjuction of any selectors. It is said that complex C covers an example e if they have the same values of attributes. For example complex $C = [A_1 = a]\ [A_3 = c]$ covers example

$e_1 = [A_1 = a]\ [A_2 = b]\ [A_3 = c]$ and does not cover

$e_2 = [A_1 = a]\ [A_2 = b]\ [A_3 = d]$.

The complex is consistent if it covers none of the negative examples.

The algorithm iteratively generates a set of consistent complexes and chooses the most preferable one. Usually such a complex is chosen, which covers as many positive examples as possible. After the selection, the positive examples covered by this complex are removed and a new iteration starts.This is continued till all positive examples are covered. The final classification rule is a disjunction of the most preferable complexes from each iteration. In order to obtain consistent complexes the following procedure called star methodology is implemented. A positive example e_i is randomly chosen. Then for each negative example f_j $j = 1, 2, \ldots, n$ a discriminant set is generated. A discriminant for e_i and f_j is the conjuction of selectors which appear in the description of e_i and not of f_j. The discriminants are generated for all the negative examples. Then by selecting one selector from each discriminant the consistent complexes which cover at least one positive example e_i are generated.

Consider a simple example. Let

$e_i = [A = a_1]\ [B = b_1]\ [C = c_1]$ be a positive example,

$f_1 = [A = a_1]\ [B = b_2]\ [C = c_1]$ and

$f_2 = [A = a_2]\ [B = b_3]\ [C = c_2]$

be all negative examples. Then a discriminantset is as follows

$d_{i1} = [B = b_1]$ and $d_{i2} = [A = a_1]\ [B = b_1]\ [C = c_1]$.

Hence the set of consistent complexes has three members:

$$C_1^i = [B = b_1]\ [A = a_1],$$
$$C_2^i = [B = b_1],$$
$$C_3^i = [B = b_1]\ [C = c_1].$$

The complex C_2^i will be chosen as the most preferable one and denoted by C^*.

Now all subsequent steps of the algorithm can be presented:

1. S_p and S_n are given sets of positive and negative examples respectively. ClassRule := Ø is a classification rule.
2. If S_p is empty, then go to 10.
3. Choose randomly a positive element $e_i \in S_p$.
4. For every element $f_j \in S_n$ generate a discriminant set.
5. Form set of consistent complexes C_m^i .
6. From C_m^i choose the most preferred one, denoted by C^*.
7. From S_p remove all examples covered by C^*.
8. ClassRule := ClassRule ∪ C^*.
9. Go to 2.
10. Stop.

It can be shown that classification rule obtained in this way is consistent and complete. It means that none of negative examples is described by this rule while all the positive ones are correctly described.

3. Modification of Michalski's algorithm

We assume that the negative examples are ordered. Also the selectors in the description of each example are ordered. Hence when we speak of *i*th negative example or *j*th attribute it is uniquely defined what we mean.

Let e_p be a given positive example from S_p. Instead of generating a discriminant set for e_p let us form zero-one matrix **D**. Matrix **D** is defined as follows.

Element d_{ji} of D is equal 1 if jth negative example has a different value than example e_p in the ith attribute. Otherwise element d_{ji} of **D** is equal 0.

Let x be a zero-one vector. The ith element of x is related to the ith attribute in the description of examples. Now let us consider the following problem:

$$\min \sum_{i=1}^{k} x_i \qquad (*)$$
$$\mathbf{D}x \geq \Lambda$$

where k is a number of attributes, Λ is a vector of ones and x_i can take value from set $\{0,1\}$.

The (*) is a set covering problem well known in integer programming. It is not difficult to see that the minimal number of variables, which cover all the rows of matrix **D** is equivalent to finding the shortest complex which covers none of the negative examples. Generally it is an NP-hard problem. Still it is a well known problem and there are many efficient heuristic algorithms giving useful solutions,Garfinkel and Nemhauser (1972).In our case we have chosen the "greedy" algorithm, because it is very simple to implement and very efficient. It can be described as follows:

An efficiency of variable x_i, $i = 1,2,\dots,k$ with respect to matrix **A** denoted $e(x_i,\mathbf{A})$ is defined as the number of ones in ith column of matrix **A**. Matrix **A** is a submatrix of **D** after deleting some rows from **D** covered in previous iterations. At the beginning **A** is equal **D**. In each iteration variable x_i^* of the greatest efficiency is selected. Then the rows covered by x_i^* are removed from **A**. The process terminates when **A** is empty, which means that all the rows of **D** are covered.

Hence in our algorithm the steps 4, 5 and 6 of original algorithm will be changed to:

4'. Form matrix **D**.

5'. Solve problem (*).

6'. Denote solution of (*) by C^*.

The other modification is of less importance with respect to avoiding combinatorial explosion. However it has a great influence on the speed of convergence of algorithm. The modification concerns the 3rd step of the algorithm. The basic idea is to choose such a positive example in each iteration which promises that its most preferable (of minimal length) consistent complex covers as many positive examples as possible. Hence the whole number of iterations should be decreased. The process of choosing the promising example is described as follows.First, the artificial/temporary object called "centroid" is formed The value of *i*th attribute in "centroid" is equal to the value of *i*th attribute which occurs most frequently in the set of positive examples. Then, such an example is chosen which is most similar to the "centroid". The measure of similarity is the number of identical selectors.

4. Results

The medical data published in Nakache and Asselain (1983) were used to test the algorithms. These data have been collected on patients with thyroid cancer, all of them were submitted to a surgical treatment. In order to get more or less the same numbers of positive and negative examples the patients were divided into classes. The positive class included patients with the survival time over 7 years. The remaining ones belonged to the negative class. Thus, 49 positive and 29 negative examples were taken into consideration. Each object was described by the following 10 attributes: sex, age, histology, metastasis, enlargement, clinical lymph nodes, clinical aspect, pathological lymph nodes, compressive syndromes and invasion. The class attribute was the survival time.

Four algorithms were applied to the data described above:

A. The algorithm as described in section 2 with some constraints as to forming large numbers of consistence complexes. The number of consistence complexes in single iteration was limited to 50.

B. The modified algorithm in which the most preferable complex was found by means of solving integer programming problem.

In A and B methods the positive example for each iteration randomly chosen.

C. The modified algorithm as in B with additional modification of selecting the most preferable complex. In single iteration the complex C^* was found for every positive example and only then the most preferable one was chosen.

D. The algorithm with both modifications described in section 3.

The results of applying the above four methods to medical data are presented and described below.

Table 1. Some parameters describing the process of finding a classification rule for the positive class.

	number of iterations	number of selectors	time
method A	11	49	2 min. 55 s
method B	11	32	52 s
method C	5	15	5 min. 38 s
method D	5	16	32 s

Table 2. Some parameters describing the process of finding a classification rule for the negative class.

	number of iterations	number of selectors	time
method A	17	60	4 min. 38 s
method B	8	14	34 s
method C	6	10	5 min. 31 s
method D	6	10	42 s

The above results are comparable, because all algorithms have been implemented in the same programming language and run on the same computer compatible with IBM

PC/AT. As can be seen the most preferable (the shortest) classification rules were obtained for methods **C** and **D**. Moreover the method D was the quickest one.

5. Summary and conclusions

In the present paper two modifications of Michalski's algorithm based on star methodology were proposed. They have influence on speed and accuracy of the algorithm (in sense of finding the most preferable rule) as well as on avoiding the combinatorial explosion without restricting the number of examples or attributes. The comparison of four algorithms shows that these modifications are essential. The modified algorithm seems to be especially useful for searching bases with large number of data.

References

[1] DIETTRICHT.G.,LONDON R., CLARKSON K., DROMEY R. Learning and inductive inference. [In:] Cohen D., Feigenbaum E. (eds.) Handbook of artificial intelligence. Kaufman, Los Altos 1981, 323-511.

[2] GARFINKEL R.S., NEMHAUSER G.L. Integer programming. J.Wiley, New York 1972.

[3] GARIS H.de Minimum length classification rules for qualitative data. Presented at the Conference in Nancy, On Applied Stochastic Modelling and Data Analysis, Dec.1988.

[4] MICHALSKI R. Discovering classification rules using variable-valued logic system VL1. Proceedings of the Third International Joint Conference on Artificial Intelligence,.IJCAI, 1973, 162-172.

[5] MICHALSKI R., CHILAUSKY R.L. Learning by being told and learning from examples: An experimental comparison of the two methods of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis. *Policy-Analysis and Information Systems*,**4** (1980) 2(June), 125-260.

[6] MICHALSKI R. A Theory and methodology of inductive learning [In:] Michalski R., Carbonell J., Mitchell T.M. (eds.), Machine Learning, Tioga, Palo Alto 1983.

[7] MICHALSKI R., STEPP R. Learning from observation: Conceptual clustering, [In:] Michalski R., Carbonell J., Mitchell T.M. (eds.), Machine Learning, Tioga, Palo Alto 1983.

[8] MITCHELL T.Generalization as search. *Artificial Intelligence*, **18** (1982), 203-226.

[9] NAKACHE J.P.,ASSELAIN B. Medical data set proposed for the workshop on data analysis. EIASM Workshop, April 1983.

[10]QUINLAN J.R. Learning efficient classification procedures and their applications to chess and games, [In:] Michalski R, Carbonell J., Mitchell T.M. (eds.) Machine Learning, Tioga, Palo Alto 1983.

[11]SHAW M.,J. Applying inductive learning to enhance knowledge-based expert systems. *Decision Support Systems*, **3**, 319-332.

Received, April 1989.

ALGORYTM UCZENIA INDUKCYJNEGO O WIELOMIANOWEJ ZŁOŻONOŚCI OBLICZENIOWEJ DO TWORZENIA NAJKRÓTSZYCH REGUŁ KLASYFIKACJI

Stosowanie w praktyce algorytmów uczenia indukcyjnego jest często nieefektywne, ze względu na zagrożenie eksplozją kombinatoryczną. Można tego uniknąć poprzez ograniczenie liczby cech (Quinlan, 1983), liczby przykładów (Garis, 1988), lub wprowadzając pewne arbitralne ograniczenia na proces uczenia indukcyjnego (Michalski, 1983). W pracy zaproponowano algorytm tworzący najkrótsze reguły klasyfikacji (w sensie liczby par cecha-wartość) o wielomianowej złożoności obliczeniowej, w którym liczba cech i/lub przykładów nie jest ograniczana. Zamieszczono wyniki obliczeń dla danych medycznych.

ИНДУКТИВНЫЕ ПРАВИЛА КЛАССИФИКАЦИИ МИНИМАЛЪНОЙ ДЛИНЫ БЕЗ УГРОЗЫ КОМБИНАТОРНОГО ВЗРЫВА

Алгоритмам индуктивного обучения угрожают резко возрастающие комбинаторные проблемы. Для избежания этой трудности рекоммендуются алгоритмы или для задач с неболъшим количеством качеств (Куинлъан, 1983), или количество примеров ограничено (Гарис, 1988), или какие то произволъные ограничения вводятся во времъя процесса обучения (Михалъски, 1983). В настоящей статъе предложен алгоритм индуктивного обучения который производит правила классификации минималъной длины (в смысле количества пар: количество-его значение). Он работает в полиномялъное время и для него количество качеств и/или примеров не ограничивается. Представлены некоторые резулътаты для медицынских примеров.