

An Inductive Learning Algorithm with a Partial Completeness and Consistence via a Modified Set Covering Problem

Janusz Kacprzyk and Grażyna Szkatuła

Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, 01-447 Warsaw, Poland
{kacprzyk, szkatulg}@ibspan.waw.pl

Abstract. We present an inductive learning algorithm that allows for a partial completeness and consistence, i.e. that derives classification rules correctly describing, e.g. most of the examples belonging to a class and not describing most of the examples not belonging to this class. The problem is represented as a modification of the set covering problem that is solved by a greedy algorithm. The approach is illustrated on some medical data.

1 Introduction

In inductive learning (from examples) we traditionally seek a classification rule satisfying *all* positive and *no* negative examples which is often strict and unrealistic. Here we assume: (1) *a partial completeness*, and (2) *a partial consistency*, i.e. that it is sufficient to describe – respectively – e.g., *most* of the positive and, e.g., *almost none* of the negative examples. We also add: (3) *convergence*, i.e. that the rule must be derived in a *finite* number of steps, and (4) that the rule is of a *minimal "length"*.

Examples are described (cf. Michalski [14]) by a set of K "attribute - value" pairs $e = \bigwedge_{j=1}^K [a_j \# v_j]$; a_j denotes attribute j with value v_j and $\#$ is a relation ($=, <, \dots$).

We propose a modified inductive learning procedure based on Michalski's [14] star-type methodology, related to our previous work (cf. Kacprzyk and Szkatuła [5 – 13]). The problem is represented as a modified set covering problem solved by a greedy algorithm. Medical data are employed for testing.

2 A Softened Problem Formulation of Inductive Learning

Sets of examples U and attributes $A = \{a_1, \dots, a_K\}$ are finite. $V_{a_j} = \{v_{i_1}^{a_j}, \dots, v_{i_j}^{a_j}\}$ is a domain of a_j , $j = 1, \dots, K$, $V = \bigcup_{j=1, \dots, K} V_{a_j}$. $f: U \times A \rightarrow V$, $f(e, a_j) \in V_{a_j}$,

$\forall a_j \in A$, $\forall e \in U$. Each $e \in U$, with K attributes, $A = \{a_1, \dots, a_K\}$, is written

$e = \bigwedge_{j=1}^K [a_j = v_i^{a_j}]$, where $v_i^{a_j} = f(e, a_j) \in V_{a_j}$ denotes attribute a_j taking on value $v_i^{a_j}$ for example e . An e in (1) is composed of K "attribute-value" pairs, denoted

$s_j = [a_j = v_i^{a_j}]$ (selectors). The conjunction of $l \leq K$ "attribute-value" pairs, i.e.

$$C^I = \bigwedge_{j \in I} s_j = \bigwedge_{j \in I} [a_j = v_i^{aj}] = [a_{j_1} = v_i^{aj_1}] \wedge \dots \wedge [a_{j_l} = v_i^{aj_l}] \tag{1}$$

where $I = \{j_1, j_2, \dots, j_l\} \subseteq \{1, \dots, K\}$ is called a *complex*.

Let us have example e and a complex $C^I = [a_{j_1} = v_i^{aj_1}] \wedge \dots \wedge [a_{j_l} = v_i^{aj_l}]$ corresponding to the set of indices $I = \{j_1, \dots, j_l\} \subseteq \{1, \dots, K\}$; $\{j_1, \dots, j_l\}$ is equivalent to a vector $x = [x_j]^T$, $j = 1, \dots, K$, such that $x_j = 1$ if $s_j = [a_j = v_i^{aj}]$ occurs in C^I , and 0 otherwise. C^I covers e if $f(C^I, a_j) = f(e, a_j), \forall j \in I$. Now, a_d is a decision attribute and $V_{a_d} = \{v_{i_1}^{a_d}, \dots, v_{i_d}^{a_d}\}$ is a domain of a_d . Each $e \in U$ is described by $\{a_1, a_2, \dots, a_K\} \cup \{a_d\}$. So, a_d determines a partition $\{Y_{v_{i_1}^{a_d}}, Y_{v_{i_2}^{a_d}}, \dots, Y_{v_{i_d}^{a_d}}\}$ of U , where $Y_{v_{i_t}^{a_d}} = \{e \in U : f(e, a_d) = v_{i_t}^{a_d}\}$, $v_{i_t}^{a_d} \in V_{a_d}$ for $t = 1, \dots, d$. Set $Y_{v_{i_t}^{a_d}}$ is called the t -th *decision class* (for $v_{i_t}^{a_d} \in V_{a_d}$), $Y_{v_{i_1}^{a_d}} \cup \dots \cup Y_{v_{i_d}^{a_d}} = U$, $Y_{v_{i_i}^{a_d}} \cap Y_{v_{i_j}^{a_d}} = \emptyset$ for $i \neq j$.

Suppose that we have a set of *positive* and *negative examples* for a class $Y_{v_{i_t}^{a_d}}$

$$S_P(Y_{v_{i_t}^{a_d}}) = \{e \in U : f(e, a_d) = v_{i_t}^{a_d}\} \tag{2}$$

$S_N(Y_{v_{i_t}^{a_d}}) = \{e \in U : f(e, a_d) \neq v_{i_t}^{a_d} \text{ and } \forall e' \in S_P(Y_{v_{i_t}^{a_d}}) \exists a_j \in P, f(e, a_j) \neq f(e', a_j)\}$
 and $S_P(Y_{v_{i_t}^{a_d}}) \cap S_N(Y_{v_{i_t}^{a_d}}) = \emptyset$ and $S_P(Y_{v_{i_t}^{a_d}}) \neq \emptyset$, $S_N(Y_{v_{i_t}^{a_d}}) \neq \emptyset$.

The descriptions of $Y_{v_{i_t}^{a_d}}$ can be given as “IF *certain conditions are fulfilled* THEN *membership in a definite class takes place*”. The rule $rul(P, v_{i_t}^{a_d})$: “IF C^I THEN $[a_d = v_{i_t}^{a_d}]$ ” is called an “*elementary*” rule for class $Y_{v_{i_t}^{a_d}}$, $v_{i_t}^{a_d} \in V_{a_d}$, $I = \{j_1, \dots, j_l\} \subseteq \{1, \dots, K\}$, $P = \{a_{j_1}, \dots, a_{j_l}\} \subseteq A \setminus \{a_d\}$, where C^I is a description of example in terms of attributes $a_j, j \in I$, and this example belongs to $Y_{v_{i_t}^{a_d}}$.

The *strength of a rule* is defined in the following manner:

$$q(rul(P, v_{i_t}^{a_d})) = \frac{card(\{e : e \in [C^I] \text{ and } f(e, a_d) = v_{i_t}^{a_d}\})}{card(\{e : e \in U\})} \tag{3}$$

We consider the classification rules:

$$\text{IF } C^{I_1} \cup \dots \cup C^{I_L} \text{ THEN } [a_d = v_i^{ad}] \tag{4}$$

with: $I_1, \dots, I_L \subseteq \{1, \dots, K\}$, $C^{I_l} = \bigwedge_{j \in I_l} [a_j = v_i^{aj}]$, $l = 1, \dots, L$.

Let us have P positive examples, $e^m \in S_P(Y_{v_i^{ad}})$, $m = 1, \dots, P$, and N negative examples, $e^n \in S_N(Y_{v_i^{ad}})$, $n = 1, \dots, N$. For each a_j , each possible value occurs at some intensity (frequency). If it occurs more frequently in the positive and less frequently in the negative examples, then it is somehow typical and should appear in the rule sought. So, we introduce the function, for each a_j , $j = 1, \dots, K$ and $v \in V_{a_j}$

$$g_j(v) = \frac{1}{P} \sum_{m=1}^P \delta(e^m, v) - \frac{1}{N} \sum_{n=1}^N \delta(e^n, v) \tag{5}$$

where: $\delta(e^m, v) = \begin{cases} 1 & \text{for } v_i^{aj} = v \\ 0 & \text{otherwise} \end{cases}$, and: $e^m \in S_P$, $v_i^{aj} = f(e^m, a_j) \in V_{a_j}$; and

analogously for $\delta(e^n, v)$. So, we may express to which degree the particular $v \in V_{a_j}$ of a_j occurs more often in the positive than negative examples; the normalized $g_j(v)$ is used as a weight of $v \in V_{a_j}$ (cf. Kacprzyk and Szkatuła [10]).

Example e_W with weights is $e_W = \bigwedge_{j=1}^K [a_j = v_i^{aj}; g_j(v_i^{aj})]$, i.e. is a conjunction of weighted selectors, $s_j^W = [a_j = v_i^{aj}; g_j(v_i^{aj})]$, that is: $C_W^I = \bigwedge_{j \in I} s_j^W$, and is called a *weighted complex*. Notice that for C_W^I x has the elements $x_j = 1$ for $j \in I$, while, for $j \in \{1, 2, \dots, K\} \setminus I$, $x_j = 0$. For C_W^I its *weighted length* is:

$$\begin{aligned} d_W(C_W^I) &= \\ &= \sum_{j \in I} (1 - g_j(v_i^{aj})) \cdot x_j + \sum_{j \in \{1, 2, \dots, K\} \setminus I} (1 - g_j(v_i^{aj})) \cdot x_j = \sum_{j=1}^K (1 - g_j(v_i^{aj})) \cdot x_j \end{aligned} \tag{6}$$

which reflects a higher relevance of those values of attributes which occur more often in the positive than in negative examples.

The length of $R_W = C_W^{I_1} \cup \dots \cup C_W^{I_L}$ is $d_{R_W}(C_W^{I_1} \cup \dots \cup C_W^{I_L}) = \max_{l=1, \dots, L} d_W(C_W^{I_l})$,

and we look for an optimal classification rule $R_W^* = C_W^{I_1^*} \cup \dots \cup C_W^{I_L^*}$ such that

$$\min_{I_1, \dots, I_L} d_{R_W}(C_W^{I_1} \cup \dots \cup C_W^{I_L}) \tag{7}$$

As the (exact) solution of (7) is very difficult, an auxiliary problem is solved (cf. Kacprzyk and Szkatuła [11]) whose solution is in general very close but much easier, i.e. an $R_W^* = C_W^{I_1^*} \cup \dots \cup C_W^{I_L^*}$ is sought such that $\min_{I_1} d_W(C_W^{I_1}), \dots, \min_{I_L} d_W(C_W^{I_L})$.

3 Formulation as a Modified Set Covering Problem

For $e^P \in S_P$, and all the negative examples $e^{P+n} \in S_N, n = 1, \dots, N$, we construct a

$$0\text{-}1 \text{ matrix } Z_{N \times K} = [z_{nj}], j = 1, \dots, K, z_{nj} = \begin{cases} 1 & \text{for } f(e^P, a_j) = f(e^{P+n}, a_j) \\ 0 & \text{for } f(e^P, a_j) \neq f(e^{P+n}, a_j) \end{cases} \text{ whose}$$

rows correspond to the consecutive $e^{P+n} \in S_N, n = 1, \dots, N$ and columns to attributes a_1, \dots, a_K ; $z_{nj} = 1$ if a_j has different values in the positive and negative examples, i.e. $f(e^P, a_j) \neq f(e^{P+n}, a_j)$, and $z_{nj} = 0$ otherwise. There are no rows with all 0s since the sets of positive and negative examples are disjoint (and non-empty). So, for any positive and negative example there is always at least one attribute with a different value in these examples.

Consider now the following inequality: $\sum_{j=1}^K z_{nj} x_j \geq \gamma_n, n = 1, \dots, N$, where

$\gamma = [\gamma_1, \dots, \gamma_N]^T$ is a 0-1 vector, and $x_j \in \{0,1\}$, for $j = 1, \dots, K$. Any vector x satisfying $Zx \geq \gamma$ determines uniquely a complex describing at least one example from the set of positive ones, and does not describe most of the negative examples. If x does not describe the n -th negative example, then $\gamma_n = 1$; and $\gamma_n = 0$ otherwise.

So, the problem is, using the above inequality:

$$\min_{x: Zx \geq \gamma} \sum_{j=1}^K (1 - g_j(v_i^{a_j})) \cdot x_j \tag{8}$$

and, in a simplified form: $\min_{x: Z^1 x \geq \gamma} d_W(C_W^{I_1}), \dots, \min_{x: Z^L x \geq \gamma} d_W(C_W^{I_L})$ in which each

minimization with respect to x is equivalent to the determination of a 0-1 vector x^* which uniquely determines the complex of the shortest weighted length. On the other hand, the satisfaction of $Zx \geq \Lambda$ (Λ is a unit vector) guarantees that such a complex would not describe all negative examples. If rules should describe *almost none* of the negative examples, the problem can be written as a modification of the set covering problem

$$\min_{x, \gamma} \sum_{j=1}^K c_j x_j \tag{9}$$

subject to: $\sum_{j=1}^K z_{1j} x_j \geq \gamma_1, \dots, \sum_{j=1}^K z_{Nj} x_j \geq \gamma_N$, with an additional constraint: $\sum_{n=1}^N \gamma_n \geq N - rel$

where $c_j = (1 - g_j(v_i^{aj}))$, $z_{nj} \in \{0,1\}$, $x_j \in \{0,1\}$, $j = 1, \dots, K$, $rel \geq 0$, $\gamma = [\gamma_1, \dots, \gamma_N]^T$, $\gamma_n \in \{0,1\}$.

This is as the original set covering problem except for that no more than rel rows are uncovered. Then, no more than rel rows can be deleted though we may loose some information, and this reduction cannot always be applied. In the set covering problem (cf. [1-4]) there is only constraint, and $\gamma = [\gamma_1, \dots, \gamma_N]^T$ is a unit vector. Problem (11) is that of covering at least $N-rel$ rows of an N -row, K -column, zero-one matrix (z_{nj}) by a subset of the columns at minimal cost c_j . We define $x_j = 1$ if column j with cost $c_j > 0$ is in the solution, and $x_j = 0$ otherwise. Then, most rows (at least $N-rel$ rows) are covered by at least one column. It always has a feasible solution (x of K element), due to the required disjointness of the sets of positive and negative examples and the way the matrix Z was constructed.

We seek a 0-1 vector x at the minimum cost and a 0-1 vector $\gamma = [\gamma_1, \dots, \gamma_N]^T$ that determines the covered rows, $\gamma_n = 1$ if row n is covered by x , and $\gamma_n = 0$, otherwise. By assumption, at least $N-rel$ rows must be covered by x . Then, an "elementary" rule for $Y_{v_i^{a_d}, v_i^{a_d}} \in V_{a_d}$, may not describe at least $(100/N \sum_{n=1}^N \gamma_n)\%$ negative examples.

The set covering problem is a well-known NP-complete combinatorial optimization problem. Many optimal and faster heuristic algorithms exist, cf. Balas and Padberg [1], Beasley [2], Christofides [3], presented a genetic algorithm, with modified operations, too. One can also use here a greedy algorithm (cf. Chvatal [4]) and we use it here.

4 An Example Using Heart Disease Data

We have 90 examples, ill or healthy, 60 are a training set and 30 are for testing. The following 12 blood factors (attributes) are measured: $lk1$ - blood viscosity for coagulation quickness 230/s, $lk2$ - blood viscosity for coagulation quickness 23/s, $lk3$ - blood viscosity for coagulation quickness 15/s, $lp1$ - plasma viscosity for coagulation quickness 230/s, $lp2$ - plasma viscosity for coagulation quickness 23/s, agr - aggregation level of red blood cells, fil - blood cells capacity to change shape, fib - fibrin level in plasma, ht - hematocrit value, sas - sial acid rate in blood serum, sak - sial acid rate in blood cells, ph - acidity of blood.

We seek classification rules into: *class 1*: patients have no coronary heart disease, *class 2*: patients have a coronary heart disease. Some results are shown below:

$A_{learning} \%$, by assumption	Number of iterations for class 1/2	Number of selectors in rule for class 1/2	$A_{learning} \%$, by assumption	$A_{learning} \%$, attained
100%	16/19	43/55	100%	90.0%
at least 97%	13/17	26/33	at least 97%	96.7%

and the results are encouraging, in fact comparable to the use of a genetic algorithm (cf. Kacprzyk and Szkatuła [13]).

5 Concluding Remarks

We proposed a improved inductive learning algorithm allowing for a partial completeness and consistence that is based on a set covering problem formulation solved by a greedy algorithm. Results seem to be very encouraging.

References

1. Balas E., Padberg M.W.: Set partitioning - A survey. In: N. Christofides (ed.) *Combinatorial Optimisation*, Wiley, New York (1979).
2. Beasley J.E.: A genetic algorithm for the set covering problem. *European Journal of Operational Research* 94 (1996) 392-404.
3. Christofides N., Korman S.: A computational survey of methods for the set covering problem. *Management Sci.* 21 (1975) 591-599.
4. Chvatal V.: A greedy heuristic for the set-covering problem. *Maths. of Oper. Res.* 4 (3) (1979) 233-235.
5. Kacprzyk J., Szkatuła G.: Machine learning from examples under errors in data, Proceedings of IPMU'1994 – 5th Int'l Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems Paris France, Vol.2 (1994) 1047-1051
6. Kacprzyk J., Szkatuła G.: Machine learning from examples under errors in data. In B. Bouchon-Meunier, R.R. Yager and L.A. Zadeh (eds.): *Fuzzy Logic and Soft Computing*, World Scientific, Singapore, (1995) 31-36.
7. Kacprzyk J., Szkatuła G.: An algorithm for learning from erroneous and incorrigible examples, *Int. J. of Intelligent Syst.* 11 (1996) 565-582.
8. Kacprzyk J., Szkatuła G.: An improved inductive learning algorithm with a preanalysis of data", In Z.W. Ras, A. Skowron (eds.): *Foundations of Intelligent Systems (Proc. of 10th ISMIS'97 Symposium, Charlotte, NC, USA)*, LNCS, Springer, Berlin (1997) 157-166.
9. Kacprzyk J., Szkatuła G.: IP1 - An Improved Inductive Learning Procedure with a Preprocessing of Data. In L. Xu, L.W. Chan, I. King and A. Fu (eds.): *Intelligent Data Engineering and Learning. Perspectives on Financial Engineering and Data Mining (Proc. of IDEAL'98, Hong Kong)*, Springer, Hong Kong (1998) 385-392, 1998.
10. Kacprzyk J., Szkatuła G.: An inductive learning algorithm with a preanalysis of data. *Int. J. of Knowledge - Based Intelligent Engineering Systems*, vol. 3 (1999) 135-146.
11. Kacprzyk J., Szkatuła G.: An integer programming approach to inductive learning using genetic algorithm. In: *Proceedings of CEC'02 - The 2002 Congress on Evolutionary Computation, CEC'02, Honolulu, Hawaii (2002)* 181-186.
12. Kacprzyk J., Szkatuła G.: An integer programming approach to inductive learning using genetic and greedy algorithms, In L.C. Jain and J. Kacprzyk (eds.): *New Learning Paradigms in Soft Computing*, Physica-Verlag, Heidelberg and New York (2002) 322-366.
13. Kacprzyk J., Szkatuła G.: A softened formulation of inductive learning and its use for coronary disease data. In: M.-S. Hacid, N.V. Murray, Z.W. Raś and S. Tsumoto (eds.) *Foundations of Intelligent Systems (Proc. of 15th ISMIS'2005 Symposium, Saratoga Springs, NY, USA)*, LNCS, Springer, Berlin, (2005), 200-209.
14. Michalski R.S.: A theory and methodology of inductive learning. In: R. Michalski, J. Carbonell and T.M. Mitchell (Eds.), *Machine Learning*. Tioga Press (1983).