Rule Discovery from Medical Data Using Genetic Algorithm

Jacek Dryl[+], Halina Kwasnicka[*], Jan Kornafel[+], Urszula Markowska-Kaczmar[*], Rafal Matkowski[+], Paweł Mikołajczyk[*], Jacek Tomasiak[*]

[+]Medical University of Wrocław

*Wroclaw University of Technology, Wyb.Wyspianskiego 27, 50-270 Wrocław, Poland

Kaczmar@ci.pwr.wroc.pl

Abstract. The paper presents a genetic algorithm used as a tool for rule discovery from real medical data. The data concern carcinoma of the cervix uteri, they cover 527 patients from three years. The last covered year is 1998, so we know results of applied therapy of patients. The efficiency of genetic algorithm used as a tool for such a problem has been studied and obtained results are shortly presented. Summary and the future work are also included in the paper.

1. Introduction

Now, with an increasing amount of data, which can be store they have became the new source of knowledge acquisition. This new knowledge can be seen as classification new patterns, discovering clusters or discovering new regularities in the data. The most popular representation describing this knowledge is expressed in the form of prepositional rules IF *premises* THEN *conclusion*. Premise is created by putting some constraints on the attribute. The conclusion standing after THEN describes class or predicted value for an attribute. While classification problem, where classes are known, seems to be easier to solve, discovering new regularity in the data is extremely difficult. First of all because we do not know whether this dependency exists, even if it exists which attributes are independent attributes and which one creates the conclusion. Next, even we have the method, which is able to discover rules with appropriate level of support in the data patterns in every case it is necessary to verify them by expert. Only he can decide whether discovered rule represents general knowledge that only confirm the appropriate workings of the method or it is very interesting rule discovering new dependency in the data. All mentioned above arguments cause that each domain problem has to be treated as unique one.

There are some methods discovering so called association rules in the data describing new regularities, but most of them are not effective, because this problem is very search intensive. One of the natural way to develop a method discovering rules is an application of genetic algorithm, which is well known technique searching a big space of possible solutions.

In this paper we present our research in designing the method based on genetic algorithm to the discovering new rules in the particular medical data (carcinoma of the cervix uteri) and our experience resulting from cooperation in the interdisciplinary group.

2. Considered medical problem

Carcinoma of the cervix uteri is the second most frequently diagnosed cancer in women in Poland. Despite of advances in diagnosis and treatment it is also one of the leading causes of cancer death. From this point of view discovering new rules formulating some new non trivial dependency in the attributes describing patients is very desirable. It could help in a better treatment.

The data are collected during 5-year observation of 527 patients with primary cancer of the cervix uteri treated in Lover Silesian Cancer Center in 1996, 1997 and 1998.

The clinicopathologic data available on these patients include:

- date of birth and patients age,
- FIGO stage of the disease (according to FIGO Staging, 1994)[3],
- tumor size,
- histological type of the tumor,
- degree of differentiation of the tumor,
- interval between diagnosis and first treatment (both dates),
- type of surgical treatment,
- type of performed radiotherapy,
- duration of radiotherapy,
- assessment of response to treatment,
- date of end of hospitalization,
- last known vital status or date of death,
- relapse-free survival,
- overall survival.

On the base of the discussion with medical experts the min aim of presented project is to discover new, nontrivial rules from examples concerning efficiency of used therapy. Exactly seeing we are interest in expected survival time of patient taking into account its state of

disease at the diagnosis moment and applied therapy. Patients that survive 5 years without recurrence are considered as cured. Our problem can be seen as strictly data mining task. As it was mentioned, to find useful rules we must search a wide space of potential rules, therefore using genetic algorithm (GA) seems to be a right attempt.

An example of using a genetic algorithm for rule discovery from a given data set we can find in the literature (e.g. [1]), but because GA is a kind of heuristic, it is necessary to develop special representation of individual, GA's operators and fitness evaluation suited for particular problem. All these details of the proposed method will be discussed in the next section

3. GA with specialized operators

In the presented problem, genetic algorithm has to find a set of nontrivial rules hidden in collected data. The two possible approaches are possible: we can evolve a population of single rules and a set of 'best rules' is obtained step by step from each generation or returned as a set of best rules from last generation (Michigan approach) or we can evolve a population of sets of rules, where one individual consists of a number of rules that together 'well' describe our data (Pitts approach). In our project we have used Michigan approach so each individual in a population represents single rule and GA returns a set of rules from the last generation.

A genetic algorithm acts as follows:

1. Generating an initial population.
2. Evaluation of individuals, storing the best individual
3. Checking stop condition: if Yes, go to step 7, if No, go to step 4.
4. Selection of individual for reproduction
5. Reproduction process: copying selected individuals to the next generation of population using mutation and crossover operators.
6. Go to step 2.
7. Stop the evolution process, return best individual from the whole evolution.

To use this general scheme of GA first of all it is necessary to determine a representation of individual, and the way of coding it in the chromosome.

3.1. Individual representation

Each individual represents a single IF-THEN rule. Generally, the rule has the following form:

IF *P1* AND *P2* THEN *LC*.

The part included in IF consists of the two kinds of premises, the first one (*P1*) refers to a medical diagnose, the second one (*P2*) to applied therapy. THEN part of a rule (a conclusion *LC*) is a probable chance for a patient to survive 5 years after finished medical treatment. Premises have typical form. They consist of a name of attribute, the comparison operator, and the value of the attribute. In our application of GA we implemented three operators (=, ≤, >) forming the following relationships:

- $Attribute_i = Value_i$,
- $Attribute_i \leq Value_i$, or
- $Attribute_i > Value_i$.

All potential attributes are divided into three classes, depending of their domains. Between them we distinguish:

- Date Attribute Class – it contains data connected with time measurement, e.g. days),
- Therapy Attribute Class – it contains data connected with used therapy, they usually are enumeration values, e.g., a kind of therapy,
- Universal Attribute Class – it may be used with different kind of attributes; they do not need special treatment.

In order to ensure the discovering rules with different premises and containing the various number of premises, evolved rules include logical flags. One flag stands before each attribute. It indicates whether the attribute is active (flag=1) or not (flag=0). All attributes values are coded using real numbers. Flags are coded in binary way, while comparison operators are coded using inter numbers. Summarizing we used so called real-coded genetic algorithm. Fig. 1 shows a part of chromosome, which codes exactly one attribute. The order of premise in the chromosome determines the name of attribute by the lookup table. All active premises in a rule are combined by AND operator.

	Flag	Comparison operator code	Value	
…	1	3	150	…

Figure 1. A part of chromosome coding a single attribute

On the base of suggestion of medical experts the conclusion of a rule is a linguistic variable in a five-point scale. It represents a chance of 5 years relapse-free survival for a patient. In the all presented rules it will be assigened by *LC*. The human expert has proposed the following scale (Table 1):

Table 1. Linguistic values for conclusion

A chance of survive	Linguistic variable
0%-20%	Bad
21%-40%	Small
41%-60%	Medium
61%-80%	Big
81%-100%	Very big

An example of rule is presented below.

IF $flag_1$ $Attrib_1 = Val_1$
AND $flag_2$ $Attrib_2 \leq Val_2$
AND $flag_3$ $Attrib_3 > Val_3$
THEN $LC = Linguistic\_Value_1$

Such a rule, is decoded from the chromosome, which creates one individual in a population. Evolving population contains the number of individuals, given by the *Population Size* parameter.

3.2. Specialized genetic operators

We have developed special genetic operators for the particular attribute class. They are modified versions of popular mutation and crossover. The overall functionality of mutation operators is the same. For each premises mutation relies on random, permissible change both: current operator of relationship, and value of considered attribute. For defined Attribute Classes mutation operators differ in the scope of changes. For all flags mutation acts in its classical form: it switches value of flag from one to zero or in opposite. The one point crossover operator acts between two selected individuals. A segmentation point is randomly selected in the individual but there is one restriction: it must lie between two attributes (solid lines in Fig. 1).

3.3. Fitness evaluation and selection

Fitness value is calculated for each individual in the whole population. It is a weighted sum of the four parts (features). The most important component has a higher weight. Below, the features considered in fitness evaluation are listed according to their importance:

- percentage coverage of the whole database by the IF part of rule decoded from the individual,
- correspondence of the rule conclusion with the appropriate data from database,
- a length of the individual,
- Correspondence of the single premise with the value of cell in the database.

Each component is scaled to the range of [0, 1] and multiplied by its weight.

For reproduction process individuals are selected according to selection method. We can use any method that fulfills one condition: it has to prefer better individuals. We use a standard roulette-wheel method [2]. Genetic operators (crossover and mutation) act on selected individuals. Crossover acts with assumed relatively high probability. Next, every attribute is mutated with assumed, small probability. The final effect of population processing creates the next generation.

4. Experiments using RuleFinder

RuleFinder is an application developed in our project, it uses genetic algorithm, described in the previous section, for rules discovering from real medical database concerning Carcinoma of the Cervix. The database is presented in [3].

4.1. Data preprocessing

The data collected in the database have to be preprocessed to the form suitable for our approach. From many columns in database experts have indicated the most significant ones. Finally, the prepared data consists of eleven columns that constitute premises data. It means that full rule can consist of 11 premises. Flag equal to zero eliminates some of the attribute from the rule during a decoding of an individual. Two columns in the processed database contain data to obtain final conclusion.

In the paper we use both names, column or attribute as alternatives. Every attribute belongs to the one of the defined class, as it was mentioned earlier. Table 2 contains the attribute classes used in the RuleFinder. Table 3 shows the names of attributes used in premises.

Table 2. The Attribute Classes in the RuleFinder

Name of class	Comparison operators	Attribute value
Date Attribute Class	$<, \leq, \geq, >$	Integer
Therapy Attribute Class	$=$	Enumeration
Universal Attribute Class	$<, \leq, =, \geq, >, \neq$	Enumeration
No-Relation Attribute Class	$=, \neq$	Enumeration

Table 3. The premise attributes in the RuleFinder

Code	Name of attribute	Class of attribute
P1	Patient age	Date Attribute Class
P2	FIGO Stage	Universal Attribute Class
P3	Tumor size	No-Relation Attribute Class
P4	Histotype	Universal Attribute Class
P5	Degree of differentiation	Universal Attribute Class
P6	Time form diagnosis to start of treatment	Date Attribute Class
P7	Surgical Code	Therapy Attribute Class
P8	RT Code	Therapy Attribute Class
P9	RT time	Date Attribute Class
P10	Response to treatment	Universal Attribute Class
P11	Time of whole treatment	Date Attribute Class

The conclusion of a rule is a very specific attribute. Only = (equal operator) can be used in the conclusion, because of it is linguistic variable with five values (see Table 1).

As it was mentioned two columns are the base to obtain the final conclusion. The first one is the time from the end of treatment to the follow-up (or death in some cases), which tells whether the patient lived five years

after the treatment or not. The second one contains information whether the patient is still alive.

3.2 Parameters of a GA

Let us remind that in the RuleFinder one individual codes one rule: the premise attributes with their flags and the specific conclusion attribute. The initial population is generated randomly. The population evolves producing better and better individuals thanks to genetic operators. But, as usually working with genetic algorithm, we must be very careful in setting the parameters, such as: the size of evolving population, probabilities of mutation and crossover, a time of evolution, etc. They are very important for assuring a balance between exploitation and exploration possibilities. The first important parameter is the number of generations. It defines the stop point for evolution. It is very hard to guess, how long the population should evolve to produce the best solution (the optimal one). After a number of experiments we have developed that 100 generation for our problem is enough. The second parameter is the size of a population. Large populations are perceived as searching large space of potential solutions and as more suitable for producing more diversified results, but, on the other hand, more different rules make a problem for experts to analyze a set of rules with relatively high fitness value. Experiments show that probabilities of mutation can be selected from the range of [0.005, 0,05] and probabilities of crossover give better results when they are assumed from the range [0.4, 0.75]. When we have well-tuned parameters concerning genetic operators – probabilities of mutation and crossover, 100 individuals in the evolved populations seems to be a reasonable size. In our project, the last set of four parameters (weights) is connected with fitness evaluation. Their values indicate percentage importance of particular components of fitness function in the final fitness evaluation of individual. A quite large number of experiments allow us to set the values as they are shown in Table 4.

Table 4. Importance of fitness evaluation features

Fitness evaluation feature	Importance for final evaluation
Percentage coverage of the whole database by evaluating individual.	60%-80%
Correspondence of the conclusion with data from database.	10%-20%
Length of the individual.	5-20%
Correspondence of the single attribute with cells from database.	5-10%

The described above values of parameters of genetic algorithm gave the best results in the initial experiments. Fig. 2 shows the changes of average fitness value of population as a function of the number of generations.

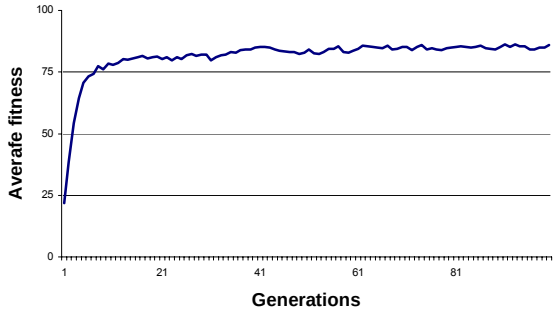

Figure 2. An average fitness of population during evolution

The *x* axis is the number of generation, which does not exceed maximum 100 generation and the *y* axis is an average value of fitness function for individuals in evolved population (scaled to the range [0, 100]. We can observe a fast improvement in the first 10 generations. Then, fitness value becomes more stable, however it grows very slowly.

4. Results of experiments

The first series of experiments were made to check whether discovered rules are true taking into account only coverage of examples in the database (the level of support). It was very important to check correctness of the fitness evaluation function. During the evaluation of individuals only the coverage of the examples is taken into account. Discovered rules are similar to those presented below:

IF P1>20 AND P4≥No biopsy AND P5≥Grade not known

THEN LC=Medium

This rule is correct for almost every patient in the database so in this way we verified that our genetic algorithm is able to find true rules. In the next step we included next components in the fitness function what should lead our method to produce more complicated and more interesting rules.

In this section we present rules obtained using RuleFinder with thoroughly selected parameters. These rules are evaluated by the experts. We focus on a few example rules. First interesting one is:

IF P1≤39
AND P6<313
AND P7=Radical abdominal hysterectomy without pelvic/paraortic lymphadenectomy
AND P8=External pelvic radiotherapy
AND P9<195
AND P10≠Stable disease
THEN LC=Medium

This rule is interesting for our experts. It says that relatively young patients, who were treated by two

mentioned by P7 and P8 therapies (see appendix) have medium chances to live five years, what is treated like the end of disease. Other attributes are connected with time of the beginning of the therapy and the time of its duration. Last attribute, P10, says that response to the treatment is positive in most cases. It is also interesting, that all attributes together do not respond to any case in the database. So, the algorithm is able to discover the rule that indirectly comes from data, but it is true and interesting for physicians.

Other interesting rule is following:

IF P2<StageIVb
AND P4<Adenosquamous carcinoma
AND P8=Interacavitary irradiation
AND P9<111
AND P10≠Progressive disease
THEN LC=Very Big.

This rule is also recognized by experts as interesting one. Attributes P2 and P4 are connected with diagnosis and attribute P10 says that response to the treatment is other than the worst. Taking into account also other premises of the rule, special radiotherapy and its time, the conclusion is that such patients have a large chance of the survival.

In our results we can find a correct rule that does not contain any information about used therapies. It says only a little about diagnosis and about duration time of the treatment. An example of such a rule is below:

IF P2≠StageIVb
AND P5<Moderately differentiated
AND P11<166
THEN LC=Big.

The most important thing concerning above example is that this rule covers almost 70% examples in the database. We can say that genetic algorithm is able to find true good rules, which are strongly connected with real data.

To our surprise, examining rules discovered by GA we noticed also rules containing in their premise part only information about diagnosis. For example:

IF P2≤StageIVb
AND P4≤Adenosquamous carcinoma
AND P5≠Moderately differentiated
THEN LC=Bad

Patients with such diagnosis expressed by the above rule have very low chance for successful treatment. This information also strongly corresponds to our real data.

The last example of discovered rule is the following:

IF P4= Adenosquamous carcinoma
AND P8=External pelvic radiotherapy
AND P9≤246
THEN LC=Big

This rule brings us information that patients with the diagnose and the therapy described in the rule, have high chance of successful treatment.

It is also worth to say that within the number of 150 rules, expert marked about 60% of them as good, possible or interesting rules.

5. Summary and future works

As it is commonly known, a genetic algorithm reveals good searching skills, but it has to be tuned for particular task. Our aim was to study possibilities of using GA as a tool for rule discovery from real, difficult for other data mining method data set. The project is realized together with physicians, they have trouble with the discovering interesting knowledge hidden in the data, using statistical methods. Our preliminary results suggest that proposed method is not a work in wrong direction. But, after these experiments we plan further developing of this project. First of all, we plan changes in data preprocessing phase. It seems, that there is strong need for preparing all data to be enumeration ones, of cause, in cooperation with our experts (physicians). This will have an affect on the used comparison operators in individuals. The *equal* (=) operator will be sufficient. In our opinion, carefully selected ranges for attributes can help in discovering rules more suitable for experts. From the expert point of view, there is no difference between, for example, patient 30 years old or 43 years old. Experts are able to provide interesting for them ranges and as a result, we hope, to obtain more interesting rules from our method. The decreasing a number of operators in the chromosome (only to '=') causes a smaller search space for genetic algorithm. It hopefully allows us to obtain good rules much easier.

Other field of potential changes is used genetic algorithm, especially a fitness function and selection method. We plan to introduce some elitism into selection method instead of simple roulette-wheel and some niching methods. They can allow to discover different optima in the search space. The tuning of the fitness function requires further experiments with different components and different weights.

The other problem is, that our approach is tested only using one medical database. In the near future we plan test it with other databases. The next will be the breast cancer database, which is collected by our experts, as well.

References

[1] Freitas A. A., *Data Mining and Knowledge Discovery with Evolutionary Algorithm.* Springer-Verlag, Berlin 2002

[2] Goldberg D. E., *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, 1989

[3] *FIGO Annual Report on the Results of Treatment in Gynecological Cancer. CARCIOMA OF THE CERVIX*. Volume 25. Patients diagnosed in 1996-1998
[4] Peña-Reyes C. A, Sipper M. Evolutionary computation in medicine: an overview. *Artificial Intelligence in Medicine* 2000; 19(1):1-23.
[5] Sadegh-Zadeh K. Fuzzy genomes. *Artificial Intelligence in Medicine* 2000; 18(1):1-28.
[6] Marvin N, Bower M, Rowe JE. An evolutionary approach to constructing prognostic models. *Artificial Intelligence in Medicine* 1999; 15(2):155-165.

Appendix A

This appendix contains a description of values of attributes used in the RuleFinder application during processing carcinoma of the cervix uteri database. Values are given in ascending order.

P1 – Integer value, which is given in years. Range from 20 to 100.

P2 – Enumeration value:
1 – Stage Ia1
2 – Stage Ia2
3 – Stage Ib1
4 – Stage Ib2
5 – Stage IIa
6 – Stage IIb
7 – Stage IIIa
8 – Stage IIIb
9 – Stage IVa
10 – Stage IVb

P3 – Enumeration value:
0 – Unknown
1 – Less or equal then 4 cm
2- Greater then 4 cm

P4 – Enumeration value:
1 – No biopsy or negative biopsy
2 – Epidermoida carcionoma
3 – Adenocarcinoma
4 – Adenosquamous carcinoma
5 – Clear cell carcinoma (mesonephroid tumor)
6 – Other (e.g. anaplastic)

P5 – Enumeration value:
0 – Grade not known/available
1 – Grade 1 – Well differentiated
2 – Grade 2 – Moderately differentiated
3 – Grade 3 – Poorly differentiated or undifferentiated carcinoma

P6 – Integer value, which is given in years. Range from 0 to 500.

P7 – Enumeration value:
0 – No surgical therapy
1 – Conization and other types of trachelectomy
2 – Simple abdominal hysterectomy without pelvic/paraortic lymphadenectomy
3 – Radical abdominal hysterectomy without pelvic/paraortic lymphadenectomy
4 – Radical abdominal hysterectomy with pelvic/paraortic lymphadenectomy

P8 – Enumeration value:
0 – No RT therapy
1 – Interacavitary irradiation (uterine and/or vaginal brachytherapy)
2 – External pelvic radiotherapy
3 – External pelvic radiotherapy + interacavitary irradiation

P9 – Integer value, which is given in years. Range from 0 to 400.

P10 – Enumeration value:
1 – CR (complete response)
2 – PR (partial response)
3 – SD (stable disease)
4 – PD (progressive disease)

P11 – Integer value, which is given in years. Range from 1 to 500.