

Rough Set Techniques for Medical Diagnosis Systems

G Ilczuk¹, R Mlynarski², A Wakulicz-Deja³, A Drzewiecka², W Kargul²

¹HEITEC AG Systemhaus fuer Automatisierung und Informationstechnologie, Erlangen, Germany

²Electrocardiology Department of Silesian Medical Academy, Katowice, Poland

³Institut of Informatics University of Silesia, Sosnowiec, Poland

Abstract

The process of discovering natural phenomena or complex system was until recently limited to finding formulas that fit empirical data. This process used with success in Science and Engineering has its limits when the complexity of the natural processes increases. Therefore, to analyze data in the medical domain an alternative approach was needed. Several mathematical methods including neural nets, inductive learning fuzzy sets and rough sets were proposed to model data in the form of decision tables or rules. Pawlak's Rough Sets Theory for handling imprecision and uncertainty in data has a main advantage over the other techniques which is the possibility to analyze data without any preliminary information what favor its usage in medical decision systems. In this paper we present the results of rule generation from our implementation of the LEM2 algorithm on reduced sets of attributes calculated using the Wrapper method with different learning algorithms.

1. Introduction

Hospital databases contain patient information that can be used by decision systems to improve medical care, uncover new relationships among data and reveal new patterns that can help to identify diseases or suggest more effective treatments. At present most research efforts in machine learning is directed toward the invention of new algorithms and much less into gaining experience in applying those to important practical applications. Therefore, in our research we focus on developing a system for doctors and clinicians which provides an insight into understanding the relations between complex medical data. The main functional blocks of the system and their functions are:

- Import the data from existing medical information systems currently in use
- Convert information extracted from narrative text into data understandable by machine learning algorithms. Information extracted in this step should include not only

knowledge directly retrieved from the free-text medical reports but also draw conclusions from the retrieved information. Information retrieved during this step will be mapped into values of several binary attributes

- Perform reduction of noisy, unreliable, irrelevant and redundant data
- Select subsets of strong attributes for further data processing
- Generate decision rules using rough set LEM2 algorithm
- Post-processing of generated decision rules such as rules pruning and joining
- Visualization of the knowledge discovery in a form easily understandable by humans

In this paper we focus on a problem of selecting subsets of strong attributes using different attribute reduction methods and analyze the classification accuracy of decision rules generated from those sets using our implementation of LEM2 algorithm.

2. Methods

2.1. Dataset preparation

The data used in our research was obtained from Electrocardiology Department of Silesian Medical Academy in Katowice-the leading Electrocardiology Department in Poland specializing in hospitalization of severe heart and arrhythmia diseases. For our experiments we took a data set of 2039 patients in average age over 70 hospitalized between 2003 and 2004. We transformed original clinical free-text reports from clinical dataset into 84 binary attributes. Each of them were then assigned to one of 14 joined attributes. The selection of joined attributes and correctness of this assignment were supervised by domain experts from the Electrocardiology Department.

Additionally information about patient's drug treatment was added to the analyzed dataset. In this paper we present experimental results for two groups of drugs commonly prescribed against heart diseases: ACE Inhibitors and Beta-blockers.

2.2. Wrapper attribute selection

The Wrapper attribute selection estimates the value of attribute subsets by using a target learning algorithm [1]. This method uses cross-validation (in our study five-fold CV) to provide an estimate for the accuracy of a classifier on the data when using only the attributes in a given subset. The CV is repeated as long as the standard deviation over the runs is greater than one percent of the mean accuracy or until five repetitions have been completed.

In this study we chose following target learning algorithms:

- Ridor (Ripple-DOWN Rule learner) has been successfully tested in medical knowledge acquisition systems and its principle to generate general rule and exceptions should simplify human analysis [2].
- JRip a rule learner proposed by Cohen as an optimized version of IREP [3].
- J48 Decision tree (C4.5 release 8) an algorithm derived from Quinlan's ID3 induction system using the standard TDIDT (top-down induction of decision trees) approach, recursively partitioning the data into smaller subsets based on the value of an attribute [4]. The "pruned" version of J48 reduces the chances of over-fitting the data [5].

For searching the space of attribute subsets we chose two different methods:

- BestFirst - which in our study always started with an empty set of attributes and searched forward by considering all possible single attribute additions.
- Genetic Search - which performs a search of attributes space using a genetic algorithm described by Goldberg [6]. For this search method we use two different population sizes: 20 (marked in result tables as GN20) and 50 (marked as GN50).

The Wrapper attribute selection method was selected over filter based methods because the interaction between the search and the learning scheme gives better results at the cost of computational expense as described by Hall [7].

2.3. Rough set preliminaries

An *information system* is a pair of $\mathbb{A} = (U, A)$, where U is a non-empty, finite set called the *universe* and A is a non-empty, finite set of *attributes*, i.e. $a : U \rightarrow V_a$ for $a \in A$, where V_a is called the *value set of a*. Elements of U are called *objects*.

Every information system $\mathbb{A} = (U, A)$ and a non-empty set $B \subseteq A$ define a *B-information function* by $Inf_B = \{(a, a(x)) : a \in B \text{ for } x \in U\}$. The set $\{Inf_A(x) : x \in U\}$ is called *A-information set* and it is denoted by

$INF(\mathbb{A})$.

We consider a special case of information system called decision tables. A decision table is any information system of the form $\mathbb{A} = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called *decision*. The elements of A are called *conditions*.

We assume that the set V_d of values of the decision d is equal to $\{1, \dots, r(d)\}$. The decision d determines the partition $\{C_1, \dots, C_{r(d)}\}$ of the universe U , where $C_k = \{x \in U : d(x) = k\}$ for $1 \leq k \leq r(d)$. The set C_k is called *k-th decision class of \mathbb{A}* .

Let $\mathbb{A} = (U, A)$ be an information system. With any subset of attribute $B \subseteq A$, an equivalence relation, denoted by $IND(B)$ called the *B-indiscernibility relation*, is associated and defined by:

$$IND(B) = \{(x, y) \in U \times U : \bigvee_{a \in B} (a(x) = a(y))\}$$

Objects x, y satisfying relation $IND(B)$ are indiscernible by attributes from B . By $[x]_{IND(B)}$ we denote the equivalence class of $IND(B)$ defined by x . A minimal subset B of A such that $IND(A) = IND(B)$ is called a *reduct* of \mathbb{A} .

If $\mathbb{A} = (U, A)$ is an information system, $B \subseteq A$ is a set of attributes and $X \subseteq U$ is a set of objects then the sets

$$\underline{B}X = \{x \in U : [x]_{IND(B)} \subseteq X\}; \text{ and}$$

$$\overline{B}X = \{x \in U : [x]_{IND(B)} \cap X \neq \emptyset\}$$

are called *B-lower* and *B-upper approximation* of X in \mathbb{A} , respectively.

2.4. LEM2 rule induction algorithm

This based on rough sets rule induction algorithm was implemented for the first time in LERS (Learning from Examples) system first induces a local covering by exploring the search space of blocks of attribute-value pairs and then converts it into the rule set [8]. To define a *local covering* a few auxiliary definitions must be presented [9].

For a variable x and its value v , a block of $[(x, v)]$ of a variable-value pair (x, v) is the set of all cases for which variable x has a value v .

Let B be a nonempty lower or upper approximation of a concept represented by a decision-value pair (d, w) . Set B depends on a set T of attribute-value pairs (a, v) if and

only if $\emptyset \neq [T] = \bigcap_{(a,v) \in T} [(a,v)] \subseteq B$.

Set T is a *minimal complex* of B if and only if B depends on T and no proper subset T' of T exists such that B depends on T' .

Let \mathbb{T} be a nonempty collection of nonempty sets of attribute-value pairs. Then \mathbb{T} is a *local covering* of B if and only if the following conditions are satisfied:

- each member T of \mathbb{T} is a minimal complex of B
- $\bigcup_{T \in \mathbb{T}} [T] = B$
- \mathbb{T} is minimal that is \mathbb{T} has the smallest possible number of members

Based on the MLEM2 algorithm description presented by Grzymala-Busse we developed a Java program to induce decision rules both from *lower approximation* (certain rules) and from *upper approximation* (possible rules)[9].

3. Results

The attribute reduction process is devised into three phases:

- reduction of unnecessary attributes
- generation of decision rules for the reduced subset of attributes and
- calculating of classification accuracy of the computed decision rules.

In this paper we used 70% of the dataset as a training set and the other 30% as a testing set. The same process of attribute reduction was performed for two decision classes - ACE and Beta-Blocker. Table 1 shows the results for the ACE decision class. From this table it can be seen that for a non-reduced dataset of 14 attributes, the total number of generated decision rules is large and the amount of time needed for the rule induction is enormous. The accuracy of classification for heavily fragmented data is also not impressive. If we compare these results with the results for a reduced subset of attributes a difference can be easily seen. Using the Wrapper attribute reduction method with different learning algorithms resulted in better classification accuracy for almost all cases. The number of generated rules was reduced 17-130 times which also resulted in a much shorter time needed for their generation. The results using J48 tree method as a learning algorithm and performing genetic search of attribute space GN50 (max generations = 50) produced very promising classification accuracy. The results show in general only small fluctuations between The number of rules, achieved classification accuracy and overall time needed for rule induction when compared with the results achieved for an unreduced set of attributes.

Table 1. Reduction of attributes for ACE drug

	Attr.	Rules	Accuracy	Time
Non-reduced set	14	536	69.99	174
W (Ridor)	6	30	75.33	3
W (JRIP)	5	12	68.96	2
W (J48+GN50)	5	18	75.48	2
W (J48+GN20)	4	13	69.10	1
W (J48)	3	4	68.76	1

Results for the Beta-blocker decision class shown in table 2 show a similar tendency. The overall lower accuracy of classification 58-59% (68-75% for the ACE) might be attributable to the increased complexity of correct prediction for this decision class. This conclusion can be confirmed with an increased number of attributes, rules and the amount of time needed for calculations. Nevertheless, also in this case, reduced sets of attributes perform more quickly and need a smaller number of rules to achieve the same or slightly better (J48,J48+GN50,JRIP) classification accuracy.

Table 2. Reduction of attributes for Beta-Blocker drug

	Attr.	Rules	Accuracy	Time
Non-reduced set	14	576	58.95	188
W (J48+GN20)	8	202	55.76	25
W (J48)	7	148	59.78	18
W (J48+GN50)	6	115	59.63	14
W (JRIP)	6	81	59.09	10
W (J48)	5	71	57.13	8

To test our implementation of LEM2 algorithm we compared the achieved results with RSES (Rough Set Exploration System) environment in version 2.2 [10]. For this comparison we generated decision rules from 70% of the entry dataset (*training set*) and used the following parameters: Method: LEM2 Algorithm, Cover parameter: 1.0. Results of the classification accuracy for the whole dataset, total number of generated decision rules and the percentage difference in comparison to our implementation are shown in table 3.

Table 3. Comparison with RSES

	Attr.	Rules	Accuracy
ACE			
Non-reduced set	14	673 (+25,56)	86,4 (+16,41)
W (Ridor)	6	55 (+83,33)	74,8 (-0,53)
W (JRIP)	5	28 (+133,33)	75,9 (+6,94)
W (J48+GN50)	5	34 (+88,89)	76,8 (+1,32)
W (J48+GN20)	4	24 (+84,62)	75,2 (+6,10)
W (J48)	3	8 (+100,00)	74,9 (+6,14)
Beta-Blocker			
Non-reduced set	14	693 (+20,31)	83,1 (+24,15)
W (J48+GN20)	8	304 (+50,50)	71,6 (+15,84)
W (J48)	7	227 (+53,38)	70,4 (+10,62)
W (J48+GN50)	6	197 (+71,30)	70,3 (+10,67)
W (JRIP)	6	146 (+80,25)	69,0 (+9,91)
W (J48)	5	126 (+77,46)	69,6 (+12,47)

From this table it can be seen that in all cases RSES system generates more decision rules and this trend is particularly visible for small sets of attributes. The classification accuracy is in all but one cases higher for RSES system but the difference for reduced sets of attributes oscillates between 6-10%. These results together with definitely better classification accuracy for non-reduced sets led us to the conclusion that our implementation of the classification system using the *bucket brigade algorithm* has some room for improvement [11]. Nevertheless, we think that these first tests of our subsystem give very positive initial results for our further work.

4. Discussion and conclusions

In this paper we have shown that attribute reduction performed as a pre-processing step before the generation of decision rules increases not only future classification accuracy but also dramatically decreases the time needed for this computation. Our experiments indicate that using the Wrapper attribute selection method together with different learning algorithms can find smaller subsets of strong attributes which performed with a similar or better accuracy during the classification. It is also worth noting that for these reduced subsets of attributes a smaller number of decision rules was needed. This criteria is especially important in the medical domain because of the need for rule testability and verification by domain experts.

Acknowledgements

We would like to thank Eugeniusz Pilat MD from the Electrocardiology Department of the Silesian Medical Academy in Katowice for giving us feedback and Maciej Pacyno MS from the statistical department of Upper Silesia Medical Center for their assistance in exporting the databases.

References

- [1] Kohavi R, John GH. Wrappers for feature subset selection. *Artificial Intelligence* 1997;97:273–324.
- [2] Preston P, Edwards G, Compton P. A 2000 rule expert system without a knowledge engineer. *Proceedings of the 8th AAAI Sponsored Banff Knowledge Acquisition for Knowledge Based Systems Workshop* 1994;17.1–17.10.
- [3] Cohen W. Fast effective rule induction. *Proceedings of the 12th International Conference on Machine Learning* 1995; ML95:115–123.
- [4] Quinlan J. C4.5, Programs for Machine Learning. San Mateo, California, USA: Morgan Kaufman, 1993.
- [5] Cohen P, Jensen D. Overfitting explained. *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics* 1997;115–122.
- [6] Goldberg DE. *Genetic Algorithms in Search, Optimization and Machine Learning*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.
- [7] Hall M, Holmes G. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on knowledge and data engineering* 2003;15:1437–1447.
- [8] Grzymala-Busse JW. *LERS-A System For Learning from Examples Based on Rough Sets*. Boston, MA: Kluwer Academic Publishers, 1992.
- [9] Grzymala-Busse JW. Mlem2-discretization during rule induction. *Proceedings of the International IIS* 2003;499–508.
- [10] Bazan JG, Szczuka MS. Rses and rseslib - a collection of tools for rough set computations. *Lecture Notes in Computer Science* 2000;2005:106–113.
- [11] Holland JH, Holyoak KJ, Nisbett RE. *Induction. Process of Inference, Learning and Discovery*. Cambridge MA London: MIT Press, 1996.

Address for correspondence:

Grzegorz Ilczuk
Am Europakanal 36/ D-91056 Erlangen / Germany
Grzegorz.Ilczuk@ilczuk.com