Rough Set Data Mining of Diabetes Data

Jaroslaw Stepaniuk

Institute of Computer Science
Bialystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
e-mail: jstepan@ii.pb.bialystok.pl

Abstract. The applications of the rough set theory to identify the most relevant attributes and to induce decision rules from a medical data set are discussed in this paper. The real life medical data set concerns children with diabetes mellitus. Three methods are considered for identification of the most relevant attributes. The first method is based on the notion of reduct and its stability. The second method is based on particular attribute significance measured by relative decrease of positive region after its removal. The third method is inspired by the wrapper approach, where the classification accuracy is used for ranking attributes. The rough set approach additionally offers the set of decision rules. For the rough set based reduced data application of nearest neighbor algorithms is also investigated. The presented methods are general and one can apply all of them to different kinds of data sets.

1 Introduction

Rough set theory was proposed [8] as a new approach to knowledge discovery from incomplete data. The rough set approach to processing of incomplete data is based on the lower and the upper approximation. The rough set is defined as the pair of two crisp sets corresponding to approximations. Some approaches to analysis of medical data sets based on the rough set theory are presented for example in [9], [10], [13], [7], [11], [12], [2], [14].

In this paper we discuss mining in diabetes mellitus data. We consider three sub-tasks:

- identification of the most relevant condition attributes,
- discovery of decision rules characterizing the dependency between values of condition attributes and decision attribute,
- application of nearest neighbor algorithms for rough set based reduced data.

The nearest neighbor paradigm provides an effective approach to classification. A major advantage of nearest neighbor algorithms is that they are non-parametric, with no assumptions imposed on the data other than the existence of a metric. However, nearest neighbor paradigm is especially susceptible to the presence of irrelevant attributes. We use the rough set approach for selection of the most relevant attributes within the diabetes data set. Next nearest neighbor algorithms are applied.

The presented approach has been applied to analyze data records of children with diabetes mellitus. This is a real life problem coming from the Second Department of Children's Diseases, Medical Academy of Bialystok, Poland.

The following features are evaluated by rough set methods and nearest neighbor algorithms on 107 patients aged 5-22 and suffering from insulin dependent diabetes for 2-13 years: sex, age of disease diagnosis, disease duration, appearance diabetes in the family, criteria of the metabolic balance, type of the applied insulin therapy, hypertension, body mass and presence or absence of microalbuminuria.

2 Description of the Clinical Data

There are two main forms of diabetes mellitus: type 1 (insulin-dependent), and the more prevalent type 2 (non-insulin-dependent). Type 1 usually occurs before age 30, although it may strike at any age. The person with this type is usually thin and needs insulin injections to live and dietary modifications to control his or her blood sugar level. Type 2 usually occurs in obese adults over age 40. It's most often treated with diet and exercise (possibly in combination with drugs that lower the blood sugar level), although treatment sometimes includes insulin therapy.

In this paper we consider data about children with insulin-dependent diabetes mellitus (type 1). Insulin-dependent diabetes mellitus is a chronic disease of the body's metabolism characterized by an inability to produce enough insulin to process carbohydrates, fat, and protein efficiently. Treatment requires injections of insulin.

Complications may happen when a person has diabetes. Some effects, such as hypoglycemia, can happen any time. Others develop when a person has had diabetes for a long time. These include damage to the retina of the eye (retinopathy), the blood vessels (angiopathy), the nervous system (neuropathy), and the kidneys (nephropathy). The typical form of diabetic nephropathy has large amounts of urine protein, hypertension, and is slowly progressive. It usually doesn't occur until after many years of diabetes, and can be delayed by tight control of the blood sugar. Usually the best lab test for early detection of diabetic nephropathy is measurement of microalbumin in the urine. If there is persistent microalbumin over several repeated tests at different times, the risk of diabetic nephropathy is higher. Normal albumin excretion is less than 20 microgram/min (less than 30 mg/day). Microalbuminuria is 20-200 microgram/min (30-300 mg/day).

Twelve condition attributes, which include the results of physical and laboratory examinations and one decision attribute (microalbuminuria) describe the database used in our experiments. The excerpt from the database is shown in Table 1. The data collection so far consists of 107 cases. The collection is growing continuously as more and more cases are analyzed and recorded. Out of twelve condition attributes eight attributes describe the results of physical examina-

Object	Sex	Age ...	Disease ...	...	HbA1c	...	Microalbuminuria
1	f	12	5	...	7.28	...	Yes
2	m	1	4	...	10.00	...	No
3	m	15	5	...	6.65	...	No
4	f	13	4	...	8.69	...	Yes
5	f	11	5	...	9.6	...	Yes
...	...	...	...	...	...	...	...
103	f	14	6	...	7.68	...	Yes
104	m	14	4	...	9.00	...	Yes
105	m	9	9	...	7.4	...	Yes
106	m	16	2	...	9.00	...	Yes
107	f	7	12	...	8.06	...	Yes

Table 1. An Excerpt of Patient Data

Symbol	Attribute	Attribute values
a_1	Sex	f, m
a_2	Age of disease diagnosis (years)	$< 7, [7,13), [13,16), \geq 16$
a_3	Disease duration (years)	$< 6, [6,11), \geq 11$
a_4	Appearance diabetes in the family	yes, no
a_5	Insulin therapy type	KIT, KIT_IIT
a_6	Respiratory system infections	yes, no
a_7	Remission	yes, no
a_8	HbA1c	$< 8, [8,10), \geq 10$
a_9	Hypertension	yes, no
a_{10}	Body mass	<3, 3-97, >97
a_{11}	Hypercholesterolemia	yes, no
a_{12}	Hypertriglyceridemia	yes, no
d	Microalbuminuria	yes, no

Table 2. Attributes and Their Values

tions, one attribute describes insulin therapy type and three attributes describe the results of laboratory examinations. The former eight attributes include sex, the age at which the disease was diagnosed and other diabetological findings. The latter three attributes include the criteria of the metabolic balance, hypercholesterolemia and hypertriglyceridemia. The decision attribute describes the presence or absence of microalbuminuria. All this information is collected during treatment of diabetes mellitus.

Additionally attributes with numeric values were discretized. Although several algorithms for automatic discretization exist (for overviews see [5]), in this analysis discretization was done manually according to medical norms. Final attributes and their values (after discretization) are presented in Table 2. Basic data information after discretization is presented in Table 3.

	%	Count
Total number of patients	100	107
Sex		
Male	54.21	58
Female	45.79	49
Age of disease diagnosis (years)		
< 7	22.43	24
$[7, 13)$	49.53	53
$[13, 16)$	22.43	24
≥ 16	5.61	6
Disease duration (years)		
< 6	51.40	55
$[6, 11)$	42.99	46
≥ 11	5.61	6
HbA1c		
< 8	42.99	46
$[8, 10)$	42.06	45
≥ 10	14.95	16
Microalbuminuria		
yes	52.34	56
no	47.66	51

Table 3. Characterization of Patients Group

3 Importance of Attributes

One can measure the importance of attributes with respect to different aspects. One can also consider different strategies searching for the most important subset of attributes. For example one can exhaust all possible subsets of the set of condition attributes and find the optimal ones. In general, its complexity (the number of subsets need to be generated) is $O\left(2^{card(A)}\right)$, where $card(A)$ is a number of attributes. This strategy is very time consuming. Therefore we consider less time consuming strategies.

In this section the importance of attributes is evaluated and compared using three methods.

3.1 Reducts Application

We compute the accuracy of approximation of decision classes. From Table 4 one can observe that both decision classes are definable by twelve condition attributes.

There are six reducts. Three reducts with nine attributes and three reducts with ten attributes. Reducts are presented in Table 5. Sign "+" means occurrence of the attribute in a reduct. Stability of reducts was verified on subtables. This idea was inspired by the concept of dynamic reducts [1]. Based on experimental verification, reducts for full data table are more stable than other attribute

Decision class	Yes	No
Number of patients	56	51
Cardinality of lower approximation	56	51
Cardinality of upper approximation	56	51
Accuracy of approximation (α)	1.0	1.0

Table 4. Accuracy of Approximation of Decision Classes

Attribute/Reduct	B1	B2	B3	B4	B5	B6
a_1	+	+	+	+	+	+
a_2	+	+	+	+	+	+
a_3	+	+	+	+	+	+
a_4	-	+	+	+	+	+
a_5	+	+	+	+	+	+
a_6	+	-	-	-	+	+
a_7	-	+	+	+	-	-
a_8	+	+	+	+	+	+
a_9	+	-	-	+	-	+
a_{10}	+	-	+	-	+	+
a_{11}	+	+	-	-	+	-
a_{12}	+	+	+	+	+	+
Stability of the reduct	65%	59%	58%	54%	47%	43%
Classification accuracy	63%	77%	71%	76%	68%	70%

Table 5. Reducts, Their Stability and Classification Accuracy

subsets. For example in one experiment we choose 30 subtables starting from 90% to 99% of all objects in data table, thus we consider 300 subtables. Six mentioned above reducts were also reducts at least in 69% from 300 subtables and other subsets were reducts in less than 10% of subtables.

In the Table 5 stability of the reducts based on four experiments is also presented. We consider 300 subtables in every experiment. The sampling strategy is the following: subtables are sampled on 10 equally spaced levels with 30 samples per level. In the following four experiments we consider different sampling levels:

Experiment 1: 60%, 64%, ..., 96% of the original table.

Experiment 2: 70%, 73%, ..., 97% of the original table.

Experiment 3: 80%, 82%, ..., 98% of the original table.

Experiment 4: 90%, 91%, ..., 99% of the original table.

In all experiments we consider subtables with at least 60% of the original table to preserve representability. On the other hand from evaluations presented in [1] we deduce that the number of at least 300 subtables is enough for good estimation of the stability coefficient.

For every reduct one can also compute classification accuracy based on leave-one-out method. The results are presented in the last row of the Table 5.

From the above analysis we infer that the reduct B2 is a relatively stable subset of attributes with high classification accuracy of generated rules.

Attribute	I	II	III	IV	V	VI
a_1	0.24	0.24	**0.15**	-	-	-
a_2	0.49	0.50	0.42	0.38	0.09	**0**
a_3	0.36	0.42	0.42	0.31	**0.09**	-
a_5	0.19	**0.19**	-	-	-	-
a_8	0.31	0.38	0.42	**0.31**	-	-
a_{12}	**0.02**	-	-	-	-	-
γ	0.76	0.74	0.55	0.40	0.09	0

Table 6. The Significance of Attributes in Succeeding Stages of the Analysis

3.2 Method Based on Significance of Attributes

For the set of all condition attributes the dependency coefficient is equal to 1.

In the first step we consider attributes that are in all six reducts. Thus we consider attributes in core. The degree of dependency is equal to 0.76. The significance of attributes is presented in Table 6. The idea is to evaluate each individual attribute with the significance measure. This evaluation results in a value attached to an attribute. Attributes are then sorted according to the values. The attribute with the least significance is removed and the process is repeated. One can stop the algorithm obtaining only one attribute.

3.3 Method Inspired by Wrapper Approach

We consider method inspired by wrapper approach [3]. The subsets of attributes are evaluated based on the cross-validation result. In the succeeding steps of the analysis attribute is removed which removal leads to the best result of the cross-validation test. The general scheme of the algorithm is as follows:

```
$B := A$;
**Repeat** $B := B - \{a\}$, where $a = \arg\max_{a \in B} \{AC(DT_{B-\{a\}})\}$.
**Until** Stop_Condition;
```

where $DT_{B-\{a\}} = (U, (B - \{a\}) \cup \{d\})$ and the resulting accuracy coefficient is $AC(DT_{B-\{a\}})$.

The partial results of the analysis are presented in Table 7. The leave-one-out test was used for accuracy estimation. The best result 79.44% was obtained for six attributes. The further removal of attributes thus not led to the increase of classification accuracy.

Every method allows to analyze data from different angle. Combining the results of the three methods one can find the following three attributes as the most important: *Age of disease diagnosis, HbA1c* and *Disease duration.* This result is consistent with the general medical knowledge about this disease.

Attribute	I	II	III	IV	V	VI
a_1	63.55	62.62	64.49	62.62	62.62	60.75
a_2	67.29	64.49	68.22	66.36	71.96	65.42
a_3	69.16	66.36	67.29	66.36	66.36	69.16
a_4	64.49	63.55	71.03	71.03	71.96	69.16
a_5	69.16	69.16	68.22	71.96	72.90	**79.44**
a_6	70.09	**72.90**	-	-	-	-
a_7	69.16	68.22	71.03	72.90	71.03	74.77
a_8	62.62	62.62	64.49	65.42	66.36	62.62
a_9	**71.03**	-	-	-	-	-
a_{10}	68.22	70.09	**76.64**	-	-	-
a_{11}	68.22	70.09	71.03	73.83	**73.83**	-
a_{12}	68.22	70.09	73.83	**74.77**	-	-
γ	1.0	1.0	1.0	0.98	0.95	0.80

Table 7. Classification Accuracy and Quality of Approximation (γ)

4 Rough Set Methods as Preprocessing for Nearest Neighbors Algorithms

In this section we discuss the experiments with nearest neighbor algorithms. The nearest neighbor algorithm retains the entire training data set during learning. This algorithm assumes all objects correspond to points in n-dimensional space. The nearest neighbors of an object are defined in terms of the Euclidean distance. More precisely, let $DT = (U, A \cup \{d\})$ be a decision table, for every two objects $x, y \in U$ the Euclidean distance is defined by $E(x,y) = \sqrt{\sum_{a \in A} (a(x) - a(y))^2}$.

Nearest neighbor algorithms are especially susceptible to the inclusion of irrelevant attributes in the data set, and several studies has shown that the classification accuracy degrades as the number of irrelevant attributes is increased (see e.g. [4]).

For number $k \in \{1, \ldots, 10\}$ of nearest neighbors and different attribute subsets (three most important attributes, all attributes and six reducts) we obtain the leave-one-out results presented in Table 8. The best results are obtained for the set $A3 = \{a_2, a_3, a_8\}$ and are also presented on Figure 1.

Conclusions

The diabetes mellitus data set has been drawn from a real life medical problem. The rough set based analysis showed that the most relevant features are the following: age of disease diagnosis, criteria of the metabolic balance and disease duration. The above aspects influence incidence of microalbuminuria in children suffering from diabetes type I. The results of our analysis and the extracted laws are also consistent with general clinical knowledge about diabetes type I. The presented methods go beyond the individual application to diabetes mellitus data analysis and can be applied to mining in different data sets.

k	$A3$	A	$B1$	$B2$	$B3$	$B4$	$B5$	$B6$
1	73.83	73.83	71.96	76.64	70.09	70.09	81.31	73.83
2	90.65	85.98	86.92	87.85	87.85	86.92	87.85	86.92
3	75.70	72.90	72.90	75.70	71.96	73.83	73.83	71.96
4	83.18	79.44	79.44	82.24	80.37	80.37	80.37	78.50
5	79.44	72.90	71.96	72.90	73.83	75.70	73.83	73.83
6	85.98	83.18	82.24	84.11	85.05	85.05	83.18	83.18
7	79.44	76.64	76.64	76.64	76.64	80.37	75.70	78.50
8	82.24	82.24	80.37	82.24	84.11	83.18	80.37	82.24
9	78.50	77.57	77.57	74.77	80.37	80.37	77.57	77.57
10	81.31	82.24	82.24	82.24	82.24	82.24	81.31	82.24

Table 8. Nearest Neighbors Method

Acknowledgments

The author would like to thank Andrzej Skowron, Jan Bazan and Dominik Slezak for valuable discussions and critical remarks. The author would like to express his gratitude to members of the Knowledge Systems Group at NTNU, Norway, and the Logic Group at Warsaw University, Poland taking part in creation of ROSETTA and to members of the Second Department of Children's Diseases, Medical Academy of Bialystok, Poland taking part in construction of the medical information system. This research was supported by the grants No. 8 T11C 023 15 and 8 T11C 010 11 from the State Committee for Scientific Research, the Bialystok University of Technology Rector's Grant W/II/3/98 and Research Program of the European Union - ESPRIT-CRIT 2 No. 20288.

References

1. Bazan J.G.: A comparison of dynamic and non–dynamic rough set methods for extracting laws from decision tables. [in:] L. Polkowski, A. Skowron, (eds.), Rough Sets in Knowledge Discovery 1. Methodology and Applications. Physica–Verlag, Heidelberg, 1998, pp. 321-365.
2. Carlin U.S., Komorowski J., Ohrn A.: Rough Set Analysis of Patients with Suspected Acute Appendicitis, Proceedings of IPMU'98, Paris, France, July 1998, pp. 1528-1533.
3. Kohavi R., John G.H.: Wrappers for Feature Subset Selection, Artificial Intelligence Journal, 97, 1997, pp. 273-324.
4. Langley P., Iba W.: Average-case Analysis of a Nearest Neighbor Algorithm, Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Mateo, CA, 1993, pp. 889-894.
5. Nguyen H.S., Nguyen S.H.: Discretization Methods in Data Mining, [in:] L. Polkowski, A. Skowron (eds.): Rough Sets in Knowledge Discovery 1. Methodology and Applications. Physica-Verlag, Heidelberg 1998, pp. 451-482.
6. Ohrn A., Komorowski J., Skowron A., Synak P.: The Design and Implementation of a Knowledge Discovery Toolkit Based on Rough Sets - The Rosetta System, [in:] L. Polkowski, A. Skowron (eds.): Rough Sets in Knowledge Discovery 1. Methodology and Applications. Physica-Verlag, Heidelberg 1998, pp. 376-399.

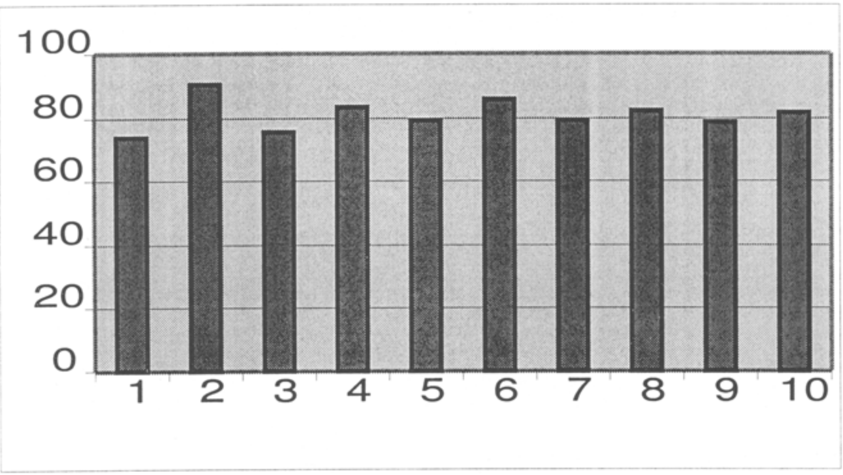

Fig. 1. Nearest Neighbors Method

7. Paszek P., Wakulicz-Deja A.: Optimization Diagnose in Progressive Encephalopathy Applying The Rough Set Theory, Proceedings of the Fourth European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany, September 2-5, 1996, vol. 1, pp. 192-196.
8. Pawlak Z.: Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht, 1991.
9. Pawlak Z., Slowinski K., Slowinski R.: Rough classification of patients after highly selective vagotomy for duodenal ulcer. International Journal of Man-Machine Studies, 24, 1998, pp. 413-433.
10. Slowinski K.: Rough Classification of HSV Patients, (ed.) Slowinski R., Intelligent Decision Support - Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic Publishers, Dordrecht, 1992, pp. 77-93.
11. Stefanowski J., Slowinski K.: Rough Set Theory and Rule Induction Techniques for Discovery of Attribute Dependencies in Medical Information Systems, Lecture Notes in Artificial Intelligence 1263, Springer-Verlag, 1997, pp. 36-46.
12. Stepaniuk J., Urban M., Baszun-Stepaniuk E.: The Application of Rough Set Based Data Mining Technique in the Prognostication of the Diabetic Nephropathy Prevalence, Proceedings of the Seventh International Workshop on Intelligent Information Systems, Malbork, Poland, June 15-19, 1998, pp. 388-391.
13. Tsumoto S., Ziarko W.: The Application of Rough Sets - Based Data Mining Technique to Differential Diagnosis of Meningoencephalitis, Proceedings of the 9th International Symposium, Foundations of Intelligent Systems, Zakopane, Poland, 9-13 June, 1996, Lecture Notes in Artificial Intelligence 1079, pp. 438-447.
14. Urban M., Baszun-Stepaniuk E., Stepaniuk J.: Application of the Rough Set Theory in the Prognostication of the Diabetic Nephropathy Prevalence. Preliminary Communication Endokrynologia, Diabetologia i Choroby Przemiany Materii Wieku Rozwojowego 1998, 4, 2.