Rakus-Andersson, Elisabeth
“Rough Sets Based on Reducts of Conditional Attributes in Medical Classification of the Diagnosis Status”
IEEE World Congress on Computational Intelligence 2008, 2008, Hong Kong, China

Rough Sets Based on Reducts of Conditional Attributes in Medical Classification of the Diagnosis Status

Elisabeth Rakus-Andersson

***Abstract*—Rough sets constitute helpful mathematical tools of the classification of objects belonging to a certain universe when dividing the universe in two collections filled with sure and possible members. In this work we adopt the rough technique to verify diagnostic decisions concerning a sample of patients whose symptoms are typical of a considered diagnosis. The objective is to extract the patients who surely suffer from the diagnosis, to indicate the patients who are free from it, and even to make decisions in undefined diagnostic cases. We also consider a decisive power of reducts being minimal collections of symptoms, which preserve the previous classification results. We use them in order to minimize a number of numerical calculations in the classification process. Finally, we test influence of symptom intensity levels on the diagnosis indisputable appearance to select these levels that are expected to be found in patients suffering from the considered diagnosis. The presence or the absence of these symptom levels in the patients allow us to add complementary remarks to earlier classification effects making them even more readable.**

I. Introduction

ROUGH set philosophy is founded on the assumption that some information is associated with every object of the considered universe set [1, 5, 6, 7, 8, 9, 10, 11, 12, 15, 16]. The objects characterized by the same information are indiscernible in view of the available information about them. The indiscernibility relation is the mathematical basis of rough set theory. Any set of indiscernible objects, being the equivalence class of the indiscernibility relation, is called an elementary set. Any union of some elementary sets (equivalence classes) is a crisp set (a precise set). Such union of elementary sets, which has boundary-line cases, i.e., objects that cannot be classified with certainty, constitutes a rough set (an imprecise, vague set).

With any rough set, a pair of precise sets – called a lower and an upper approximation of the rough set – is associated. The lower approximation consists of all objects that are surely included in the set, and the upper approximation contains all objects that definitely or possibly belong to the set. A difference between the upper and the lower approximation constitutes the boundary region of the rough set. Approximations are two basic operations in the rough set theory.

This work was supported in part by the Swedish Royal Academy of Sciences under Grant in the Academic Class Mathematics, 2007.

E. Rakus-Andersson is with Blekinge Institute of Technology, School of Engineering, Department of Mathematics and Science, S-37179 Karlskrona, Sweden (phone: +46455385408; fax: +46455385460; e-mail: Elisabeth.Andersson@bth.se).

We utilize this technique to classify some patients who are supposed to suffer from the same diagnosis. The presence of the diagnosis is primarily confirmed, denied or undefined in the patients. We intend to verify the predetermined hypotheses by stating sets of patients assigned to classes of the considered diagnosis surely or possibly and, by the way, we want to find patients who are attached to a class that does not confirm the diagnosis presence.

In Section II we first discuss the theoretical aspect of adaptation of rough set axioms to the medical task to prove the clues in a clinical exercise afterwards. We also would like to reduce a number of clinical data concerning symptoms, treated here as conditional attributes of the model, without depriving the data of their decisive character. A selection of minimal samplings of symptoms possessing full power in decision-making is accomplished in Section III.

Since we still intend to improve classification results obtained for reducts then we will consider influence of differentiated symptom levels on the sure diagnosis occurrence. The presence, the absence or the partial presence of selected levels in patients provides us with additional comments clearing the patients' diagnostic status. The last investigations constitute the item of Section IV. In Section V we sum up achievements of this work.

II. Rough Set Theory in the Diagnostic Classification of Patients

Let us first introduce the theoretical background of rough sets and let us then prove their usefulness via presenting a practical problem concerning medical diagnosing. All conceptions and annotations will be accommodated to a medical model to make it ready for the practical interpretation.

A. Theoretical Assumptions of the Classification Model

We start with the information system constructed as a data decision table whose columns are labeled by attributes. Objects of interest label the table rows, and entries of the table are attribute values.

In a scenario of the diagnostic discussion already sketched in the introduction and interpreted as a classification of objects assigned to a certain diagnosis, we adopt the set of patients $P = \{P_1, \ldots, P_m\}$ possessing objects P_i, $i = 1, \ldots, m$, as a *universe set* P. The set of *condition attributes* S is established as a set of symptoms $S = \{S_1, \ldots, S_n\}$ [8, 13, 14]. With every attribute $S_j \in S$, $j = 1, \ldots, n$, we associate a set

$V_{S_j} = \{x^1_{S_j}, x^2_{S_j}, ..., x^{t(S_j)}_{S_j}\}$ of its values, called the *domain* of S_j. In the diagnostic problem the set V_{S_j} will contain either linguistic terms or values of the membership degrees of S_j that are, in both cases, expressed by codes corresponding to the intensity grades of S_j. Any subset $B = \{S_{j_1}, ..., S_{j_p}\}$, $p \leq n$, of S, consisting of some selected symptoms among S_1, …, S_n, determines a binary relation $I(B)$ on P, which will be called an *indiscernibility* relation. The relation $I(B)$ is found as a set of pairs

$$(P_i, P_l) \in I(B) \text{ if } S_{j_k}(P_i) = S_{j_k}(P_l) \tag{1}$$

for each $S_{j_k} \in B \subseteq S$, $i, l = 1, \ldots, m$, $j = 1, \ldots, n$, $k = 1, \ldots, p$, where $S_{j_k}(P_i)$ denotes the value $x^c_{S_{j_k}}$, $c = 1, \ldots, t(S_{j_k})$, of attribute S_{j_k} for the element P_i.

The relation $I(B)$, as reflexive, symmetric and transitive, is recognized as an equivalence relation.

We make a partition of the set P, with respect to B, by means of the relation $I(B)$ to obtain equivalence classes $IB(P_i)$ defined by

$$IB(P_i) = \{P_l : (P_i, P_l) \in I(B)\} \tag{2}$$

for each $i, l = 1, \ldots, m$.

The classes $IB(P_i)$ are additionally called elementary sets. We realize that these sets contain the objects P_i that are identical, i.e., in the considered case, they sample patients who suffer from presence of the same symptoms characterized by the same intensity.

The symptoms S_{j_k}, $k = 1, \ldots, p$, constitute the condition attributes in the diagnostic model of classification. Besides these, we also consider a decision attribute – the diagnosis D_1 – that is initially recognized with a different status in the patients from set P. D_1 has a set of values determined as "*yes*" if it has been found in the patient, "*no*" if the patient seems to be free from it and "*unknown*" when a decision about the presence of the diagnosis cannot be clearly formulated.

By resuming the assumptions made so far we can come to a conclusion that the contents of the classification table, giving rise to the indiscernibility relation $I(B)$, corresponds to a *triple* (P, B, D_1) in the model of the diagnosis assignment to each considered patient. The patients P_i are placed in the first column of the table; the three values of D_1 appear in the last column whereas the rest of the table positions are filled with the values of condition attributes S_{j_k}, i.e., codes assigned to S_{j_k}.

The aim of the classification, accomplished by $I(B)$ or rather its equivalence classes, is to divide the patients belonging to P in three groups. These three groups are: a group of patients who surely are ill with D_1, a sample of patients who suffer or may suffer from D_1 and a collection of patients who do not have diagnosis D_1.

Let us create a set $P_{yes} \subseteq P$ in accordance with the following definition

$$P_{yes} = \{P_i : D_1 \text{ has decision "}yes\text{" assigned in the table}\} \tag{3}$$

for $i = 1, \ldots, m$.

We now state two sets that surround $P_{yes} \subseteq P$. These sets constitute P_{yes}'s lower and upper approximations.

The lower approximation $B_*(P_{yes})$ of P_{yes} is built as

$$B_*(P_{yes}) = \{P_i : IB(P_i) \subseteq P_{yes}\} \tag{4}$$

and is apprehended to be a set of these P_i, who have D_1 assigned with a full security.

The other set, the upper approximation $B^*(P_{yes})$ of P_{yes}, is designed by

$$B^*(P_{yes}) = \{P_i : IB(P_i) \cap P_{yes} \neq 0\} \tag{5}$$

and is accepted as a sampling of those objects P_i that surely or possibly are members of the D_1-class possessing the attribute "*yes*" (D_1 = "*yes*").

The set P_{yes} is thus bounded by two sets in compliance with the inclusion $B_*(P_{yes}) \subseteq P_{yes} \subseteq B^*(P_{yes})$ and referred to the approximation sets as rough or inexact with respect to B.

Even a boundary region of P_{yes}, denoted by $B_{bn}(P_{yes})$ and equal to

$$B_{bn}(P_{yes}) = B^*(P_{yes}) - B_*(P_{yes}) \tag{6}$$

contains some useful information about attendance of the objects that are uncertain members of the class D_1 = "*yes*".

To measure a grade of membership uncertainty in the D_1 = "*yes*" class for each P_i, we apply a formula

$$\mu_{D_1 = "yes"}(P_i) = \frac{|P_{yes} \cap IB(P_i)|}{|IB(P_i)|}. \tag{7}$$

A selection of the B-subset of S should be made with the special care to assure good classification results, i.e., we wish to avoid making too great differences between the contents and cardinalities of approximating sets. We can measure a coefficient α_B called *the accuracy of approximation* in conformity with

$$\alpha_B(P_{yes}) = \frac{|B_*(P_{yes})|}{|B^*(P_{yes})|} \tag{8}$$

to state the grade of roughness of the set P_{yes}.

B. The practical Explanation of the Patients' Allocation within Diagnostic Classes

We demonstrate the utility of rough sets in the diagnosis classification process by studying steps of the following example.

Example 1

A physician has listed 10 symptoms that are the elements of the set of symptoms $S = \{S_1 -$ "*hereditary inclination*", $S_2 -$ "*ECG changes in resting position*", $S_3 -$ "*smoking*", $S_4 -$ "*lack of physical activity*", $S_5 -$ "*pain in chest*", $S_6 -$ "*breathlessness*", $S_7 -$ "*feeling of sickness*", $S_8 -$ "*hypertension*", $S_9 -$ "*increased level of LDL-cholesterol*", $S_{10} -$ "*obesity*"$\}$. These are associated with three diagnoses $D_1 =$ "*high risk of cardiovascular diseases*", $D_2 =$ "*coronary heart disease*" and $D_3 =$ "*myocardial infarct*".

Let us select set $B \subseteq S$ as $B = \{S_3, S_4, S_8, S_9, S_{10}\}$. Set B contains the most significant symptoms for diagnosis D_1.

We now prepare sets of values describing intensities of the selected symptoms.

The symptoms S_3 and S_4 are compound qualitative parameters measured by means of a questionnaire while S_8, S_9 and S_{10} are the quantitative indicators. By using the adaptive techniques for biological parameters S_3, S_4, S_8, S_9 and S_{10} to convert them to fuzzy sets S_j, $j = 3, 4, 8, 9, 10$, with corresponding membership degrees $\mu_{S_j}(P_i)$ [12, 13] we furnish the symptoms with numerical representatives coming from the continuous interval [0, 1]. The property of expressing the symptoms' intensity grades over the same interval [0, 1] makes the considered parameters comparable in spite of their different characteristics. Further, in order to vary the grades as discrete characteristic quantities, we construct the following codes associated with the membership values $\mu_{S_j}(P_i)$, $j = 3, 4, 8, 9, 10$, belonging to subintervals of [0, 1]. We assign the code 0 to $\mu_{S_j}(P_i) \in [0, 0.25)$, 1 – to $\mu_{S_j}(P_i) \in [0.25, 0.5)$, 2 – to $\mu_{S_j}(P_i) \in [0.5, 0.75)$ and, finally, 3 – to $\mu_{S_j}(P_i) \in [0.75, 1]$. The codes generate sets $V_{S_j} = \{0,1,2,3\}$, $j = 3, 4, 8, 9, 10$.

Assume that $P = \{P_1, P_2, P_3, P_4, P_5, P_6\}$. The patients P_1, P_2 and P_5 are supposed to suffer from D_1, P_3 and P_6 have D_2 assigned, and the diagnosis concerning P_4 is unknown. We decide the members of set $P_{yes} = \{P_1, P_2, P_5\}$. To regard P_{yes} as rough, we should find its lower and upper approximation. In this way we also count on classifying the unknown object P_4.

We now fill the entries of TABLE I, known as (P, B, D_1), that constitutes a basis for establishing an indiscernibility relation $I(B)$.

TABLE I
(P, B, D_1) IN DIAGNOSIS CLASSIFICATION

Patients	Codes characteristic of symptoms					Decision about D_1
	S_3	S_4	S_8	S_9	S_{10}	
P_1	1	3	2	1	2	*yes*
P_2	2	3	3	2	1	*yes*
P_3	0	2	1	1	1	*no*
P_4	2	3	3	2	1	*unknown*
P_5	3	3	2	2	2	*yes*
P_6	0	1	1	2	3	*no*

The relation $I(B)$ consists of the pairs (P_i, P_l), $i, l = 1, \ldots, 6$, containing patients who, when comparing rows i and l, have all symptom codes equal.

We list $I(B)$ as $I(B) = \{(P_1, P_1), (P_2, P_2), (P_3, P_3), (P_4, P_4), (P_5, P_5), (P_6, P_6), (P_2, P_4), (P_4, P_2)\}$.

The elementary sets of $I(B)$ or its equivalence classes are given as the sets $IB(P_1) = \{P_1\}$, $IB(P_2) = \{P_2, P_4\}$, $IB(P_3) = \{P_3\}$, $IB(P_4) = \{P_2, P_4\}$, $IB(P_5) = \{P_5\}$, $IB(P_6) = \{P_6\}$.

The lower approximation of P_{yes} is established as the set $B_*(P_{yes}) = \{P_1, P_5\}$ while P_{yes}'s upper approximation is obtained as $B^*(P_{yes}) = \{P_1, P_2, P_4, P_5\}$.

The boundary region of P_{yes} is found as the set $B_{bn}(P_{yes}) = \{P_2, P_4\}$.

The sizes of the membership degrees, confirming the patients' attendance in the $D_1 =$ "*yes*" class, have been evaluated as

$$\mu_{D_1="yes"}(P_1) = 1, \quad \mu_{D_1="yes"}(P_2) = 1/2, \quad \mu_{D_1="yes"}(P_3) = 0,$$
$$\mu_{D_1="yes"}(P_4) = 1/2, \quad \mu_{D_1="yes"}(P_5) = 1, \quad \mu_{D_1="yes"}(P_6) = 0.$$

We can assume that P_1 and P_5 have D_1 with a one hundred percent confidence, while P_2 and P_4 may suffer from D_1. We can also notice that P_4 affects a status of P_2 negatively, and on the contrary, we can see that P_2 upgrades an importance of P_4 as a member in the $D_1 =$ "*yes*"-class.

The accuracy approximation coefficient $\alpha_B(P_{yes}) = 1/2$ measures the grade of imprecision of the set P_{yes} in the meaning of its roughness when comparing to a crisp set.

III. THE SELECTION OF REDUCTS FROM A SET OF CONDITIONAL ATTRIBUTES

The indiscernibility relation reduces the data by identifying the equivalence classes since only one element of the equivalence class is entailed to represent the entire class.

On the other hand, we sometimes observe the presence of superfluous data brought in the decision table (P, B, D_1) by some needless attributes belonging to $B \subset S$. To remove the unnecessary conditional attributes from S (or, particularly, from its subset B) and, at the same time, to preserve the induction of the same approximation sets of P_{yes} we try to extract collections of B's subsets being *minimal* sets called *reducts*. Their applications as conditional attributes warrant that the essence of information obtained earlier will be invariable [1, 2, 3, 4, 5, 6, 15, 17, 18, 19].

A. The Generation of Reducts by Dependency Rules

Let us still consider the set of patients $P = \{P_1, \ldots, P_m\}$ and the set of symptoms $S = \{S_1, \ldots, S_n\}$. For each symptom S_j, $j = 1, \ldots, n$, we adopt the set $V_{S_j} = \{x^1_{S_j}, x^2_{S_j}, \ldots, x^{t(S_j)}_{S_j}\}$ of its values. The status of diagnosis D_1 is still recognized as a value of the decision attribute.

For any subset $B = \{S_{j_1}, \ldots, S_{j_p}\}$, $p \le n$, of S we can determine an $p \times p$ discernibility relation $M^P_{D_1}(B)$ of pairs (P_i, P_l), $i, l = 1, \ldots, m$, with associated entries e_{il} introduced by the definition

$$e_{il} = \{S_{j_k} \in B : S_{j_k}(P_i) \neq S_{j_k}(P_l)\}, \tag{9}$$

$k = 1, \ldots, p$.

Hence, in each cell of matrix $M_{D_1}^{P}(B)$ we sample these symptoms that take different values of codes for two compared patients.

Further, a discernibility function $f_{D_1}^{P}(B)$ is a function defined by

$$f_{D_1}^{P}(B) = \wedge\{\vee(e_{il}) : 1 \le i, l \le m, i < l, e_{ij} \ne 0\}. \quad (10)$$

where $\vee(e_{il})$ is the logical disjunction of the symptoms $S_{j_k} \in e_{il}$, while $\wedge$ stands for the logical conjunction of listed disjunctions.

A dependency rule is a disjunctive normal form of (10). This emerges groups of symptoms belonging to B. These should maintain the previous results obtained for the entire B. It means that we expect to get the same lower and upper approximation of set P_{yes} even if the approximation is generated by reduced collections of symptoms.

Let us add a decision D_1 to sets P and B in the further development of the *reduct information system*. D_1 is characterized by the set of values already mentioned as d_1 = "*yes*", d_2 = "*no*" and d_3 = "*unknown*". The set $V_{D_1} = \{d_1, d_2, d_3\}$ represents three different decision classes. We now divide the decision table (P, B, D_1) into three tables (P^b, B, $D_1 = d_b$), $b = 1, 2, 3$, due to three decision states d_1, d_2, d_3.

Let us note that, within each class, we refer to the set of patients $P^b = \{P_{i_1}^b, ..., P_{i_h}^b\}$, $h \le m$, gathering the objects of P associated with the decision state d_b, $b = 1, 2, 3$. For each d_b-decision we determine a discernibility matrix $M_{d_b}^{P^b}(B)$ with the entries $e_{i_x l_y}$ in conformity with

$$e_{i_x l_y} = \{S_{j_k} \in B : S_{j_k}(P_{i_x}^b) \ne S_{j_k}(P_{l_y}^b)\} \quad (11)$$

for $x, y = 1, \ldots, h$.

For the sets P^b and the decisions d_b the discernibility function $f_{d_b}^{P^b}(B)$ is defined as

$$f_{d_b}^{P^b}(B) = \wedge\{\vee(e_{i_x l_y}) : 1 \le i_x, l_y \le m, \\ 1 \le x, y \le h \le m, i_x < l_y, e_{i_x j_y} \ne 0\} \quad (12)$$

where $\vee(e_{i_x l_y})$ is a disjunction of all members of $e_{i_x l_y}$. Afterwards, by using logical laws for conjunctions and disjunctions (especially we can mention the commutative, associative, distributive and absorption laws) we convert $f_{d_b}^{P^b}(B)$ to its disjunctive normal form (*d.n.f*) known as a dependency rule for obtaining reducts. The disjunction now will be an outer function tying together the brackets having conjunction of symptoms as an inner function. The contents of each bracket, in which the symptoms S_{j_k} are joined by conjunction $\wedge$, provides us with a new sample of symptoms that are the conditional attributes of a decision table assimilated only to them. The table (P, *a reduct of* B, D_1), in turn, collects fundamental data to build an indiscernibility relation that should give us the original approximation sets determined for set B. In this way we exclude symptoms of less importance for the classification of D_1 without making changes in final classification results.

B. Applications of Reducts to Selections of D_1's Classes

We return to Ex. 1 to use its data in the further investigations concerning the classification of patients on the basis of reducts.

Example 2

We recall set $B \subseteq S$ stated as $B = \{S_3, S_4, S_8, S_9, S_{10}\}$. In conformity with three values of the decision attribute D_1 equal to d_1 = "*yes*", d_2 = "*no*" and d_3 = "*unknown*" we intend to split TABLE I in three tables.

Let us only create a table associated with d_1 = "*yes*" because the acceptance of D_1 is the most significant decision in the considered classification. We thus reorganize TABLE I as TABLE II by deleting all rows of TABLE I that are not marked by the decision "*yes*". The set $P^1 = \{P_1, P_2, P_5\}$.

TABLE II
THE TABLE (P^1, B, $D_1 = d_1$)

Patients	Codes characteristic of symptoms					$D_1 = d_1$
	S_3	S_4	S_8	S_9	S_{10}	
P_1	1	3	2	1	2	*yes*
P_2	2	3	3	2	1	*yes*
P_5	3	3	2	2	2	*yes*

The discernibility matrix $M_{d_1}^{P^1}(B)$, determined on the basis of TABLE II for pairs of P_1, P_2 and P_5 respectively due to (11), contains entries filled with the symptoms whose values differ from each other for two compared patients. Thus, $M_{d_1}^{P^1}(B)$ takes a form of TABLE III.

TABLE III
THE MATRIX $M_{d_1}^{P^1}(B)$

Patients/Patients	P_1	P_2	P_5
P_1		S_3, S_8, S_9, S_{10}	S_3, S_9
P_2			S_3, S_8, S_{10}
P_5			

In accordance with (12) we derive the formula of the discernibility function $f_{d_1}^{P^1}(B)$ as

$$f_{d_1}^{P^1}(B) = (S_3 \vee S_8 \vee S_9 \vee S_{10}) \wedge (S_3 \vee S_9) \wedge (S_3 \vee S_8 \vee S_{10}).$$

We adopt the logical laws for the conjunction and the disjunction to expand $f_{d_1}^{P^1}(B)$ in the *d.n.f.*-form. Hence

$$f_{d_1}^{P^1}(B) = (S_3 \vee S_8 \vee S_9 \vee S_{10}) \wedge (S_3 \vee S_9) \wedge (S_3 \vee S_8 \vee S_{10})$$
$$= [[(S_3 \vee S_9) \vee (S_8 \vee S_{10})] \wedge (S_3 \vee S_9)] \wedge (S_3 \vee S_8 \vee S_{10})$$
$$= (S_3 \vee S_9) \wedge [S_3 \vee (S_8 \vee S_{10})] = S_3 \vee [S_9 \wedge (S_8 \vee S_{10})]$$
$$= S_3 \vee (S_8 \wedge S_9) \vee (S_9 \wedge S_{10}).$$

The disjunctive normal form of $f_{d_1}^{P^1}(B)$ generates reducts of set B, i.e., the sets $B_{d_1}^1 = \{S_3\}$, $B_{d_1}^2 = \{S_8, S_9\}$ and $B_{d_1}^3 = \{S_9, S_{10}\}$. These configurations of symptoms should replace the contents of set B in new decision tables without inserting different information about the rough set P_{yes}.

Let us prove set $B_{d_1}^1 = \{S_3\}$ as a set of new condition attributes. Then, TABLE IV yielding a new view of TABLE I is stated for only S_3.

TABLE IV
(P, $B_{d_1}^1$, D_1) IN DIAGNOSIS CLASSIFICATION

Patients	Codes characteristic of symptoms					Decision about D_1
	S_3	deleted data	deleted data	deleted data	deleted data	
P_1	1					*yes*
P_2	2					*yes*
P_3	0					*no*
P_4	2					*unknown*
P_5	3					*yes*
P_6	0					*no*

$I(B_{d_1}^1)$ = {(P_1, P_1), (P_2, P_2), (P_3, P_3), (P_4, P_4), (P_5, P_5), (P_6, P_6), (P_2, P_4), (P_4, P_2), (P_3, P_6), (P_6, P_3)}. The elementary sets of $I(B_{d_1}^1)$ are determined as $IB_{d_1}^1(P_1) = \{P_1\}$, $IB_{d_1}^1(P_2) = \{P_2, P_4\}$, $IB_{d_1}^1(P_3) = \{P_3, P_6\}$, $IB_{d_1}^1(P_4) = \{P_2, P_4\}$, $IB_{d_1}^1(P_5) = \{P_5\}$, $IB_{d_1}^1(P_6) = \{P_3, P_6\}$.

The lower approximation of $P_{yes} = \{P_1, P_2, P_5\}$ is equal to $B_{d_1*}^1(P_{yes}) = \{P_1, P_5\}$ and its upper approximation is a set $B_{d_1}^{1\,*}(P_{yes}) = \{P_1, P_2, P_4, P_5\}$. The approximation sets surrounding P_{yes} are exactly the same as obtained by means of B.

We test the next set of symptoms $B_{d_1}^2 = \{S_8, S_9\}$ in the classification of D_1. TABLE V contains the rearranged data influenced by new condition attributes.

TABLE V
(P, $B_{d_1}^2$, D_1) IN DIAGNOSIS CLASSIFICATION

Patients	Codes characteristic of symptoms					Decision about D_1
	deleted data	deleted data	S_8	S_9	deleted data	
P_1			2	1		*yes*
P_2			3	2		*yes*
P_3			1	1		*no*
P_4			3	2		*unknown*
P_5			2	2		*yes*
P_6			1	2		*no*

Since $I(B_{d_1}^2)$ = {(P_1, P_1), (P_2, P_2), (P_3, P_3), (P_4, P_4), (P_5, P_5), (P_6, P_6), (P_2, P_4), (P_4, P_2)} is exactly the same as $I(B)$ from Ex. 1 then the partition of P and the approximating sets are not expected to change. This confirms that the application of set $B_{d_1}^2$, truncated when comparing to B, preserves the effects brought by B and reduces the number of performed operations in the process of the attribute comparison.

At last we set $B_{d_1}^3 = \{S_9, S_{10}\}$ as condition attributes of the classification. The modified TABLE I appears as TABLE VI.

TABLE VI
(P, $B_{d_1}^3$, D_1) IN DIAGNOSIS CLASSIFICATION

Patients	Codes characteristic of symptoms					Decision about D_1
	deleted data	deleted data	deleted data	S_9	S_{10}	
P_1				1	2	*yes*
P_2				2	1	*yes*
P_3				1	1	*no*
P_4				2	1	*unknown*
P_5				2	2	*yes*
P_6				2	3	*no*

Even though we have cut off some symptoms from B, the indiscernibility relation $I(B_{d_1}^3)$ = {(P_1, P_1), (P_2, P_2), (P_3, P_3), (P_4, P_4), (P_5, P_5), (P_6, P_6), (P_2, P_4), (P_4, P_2)} decided for $B_{d_1}^3 = \{S_9, S_{10}\}$ is still invariable when comparing to B's results. This means that P_{yes} will be located as rough in the same neighborhood of two approximation sets.

IV. CLASSIFICATION OF DISEASES BASED ON IMPORTANCE LEVELS OF REDUCT MEMBERS

The presence of values d_1 = "*yes*", d_2 = "*no*" and d_3 = "*unknown*", that constitute the contents of the last columns in the decision tables established before, has been decided by a physician according to his experience. We wish to verify his judgment concerning the status of D_1 in patients by proposing a classification of levels differentiating intensities of symptom occurrences. We consider the symptoms that are members of any reduct selected in the previous subsection.

By proposing a new method involved in a creation of another rough set we intend to select the symptom grades constituting the most severe threat for the patients' condition in regard to D_1.

Let us recall reduct $B_{d_1}^2 = \{S_8, S_9\}$, whose two symptoms S_8 and S_9 can be, due to the code values 0, 1, 2 and 3, split in intensity levels "*low*" = "*l*", "*medium*" = "*m*" and "*high*" = "*h*". We thus introduce new objects $S_8^{"l"}$ for $V_{S_8}^{"l"} = \{0,1\}$, $S_8^{"m"}$ for $V_{S_8}^{"m"} = \{2\}$ and $S_8^{"h"}$ for $V_{S_8}^{"h"} = \{3\}$. Symptom S_9 is represented by levels $S_9^{"l"}$ for $V_{S_9}^{"l"} = \{0\}$, $S_9^{"m"}$ for $V_{S_9}^{"m"} = \{1,2\}$ and $S_9^{"h"}$ for $V_{S_9}^{"h"} = \{3\}$ after preparing another

code rule. Let us collect all standards of S_8 and S_9 in a set $S_{B^2_{d_1}} = \{S_8^{"l"}, S_8^{"m"}, S_8^{"h"}, S_9^{"l"}, S_9^{"m"}, S_9^{"h"}\}$.

Further we remind of the existence of the rough set $P_{yes} = \{P_1, P_2, P_5\}$ surrounded by the lower approximation $B^2_{d_1*}(P_{yes}) = \{P_1, P_5\}$ and the upper approximation $B^{2\,*}_{d_1}(P_{yes}) = \{P_1, P_2, P_4, P_5\}$ determined on the basis of the reduct $B^2_{d_1} = \{S_8, S_9\}$. We thus suggest that different stages of S_8 and S_9 should be analyzed in patients P_1 and P_5 who are pointed out by $B^2_{d_1*}(P_{yes})$ as the objects in which D_1 is surely located. From the medical point of view we assume that only the levels "*medium*" and "*high*" are particularly essential for the confirmation of D_1's presence in patients.

To verify the hypothesis about the special significance of levels "*medium*" and "*high*", expressed above, we build a new decision table ($S_{B^2_{d_1}}$, $B^2_{d_1*}(P_{yes})$, risk of symptom level for D_1). In the table we make another match of attributes, namely, the objects are interpreted as symptom levels, the condition attributes are now the patients selected by the lower approximation set of P_{yes} obtained for S_8 and S_9 and, finally, the decision attribute is defined as "risk of symptom level for D_1" To explain the last statement we add that we form two states of the risk. We place "*yes*" in the last position of the decision table row if any symptom level $S_8^{"m"}$, $S_8^{"h"}$, $S_9^{"m"}$ or $S_9^{"h"}$ is met in the first position of the same row, whereas "*no*" stands for the levels $S_8^{"l"}$, $S_9^{"l"}$ typical of the absence of D_1 when believing in the physician's expertise. To fill in the part concerning information about conditional attributes we should, on the basis of the data placed in TABLE I, assign "+" to P_1 and P_5 if the symptom levels from set $S_{B^2_{d_1}}$, due to their definitions involving code values, are found in the patients. Opposite, the sign of "–" emphasizes that P_1 and P_5 are free from the levels indicated by the first column of the table.

By making these preparations we count on finding of the lower and upper approximation sets of $(S_{B^2_{d_1}})_{yes}$ = {set of symptom levels that have decision "*yes*" assigned in the last column of the decision table ($S_{B^2_{d_1}}$, $B^2_{d_1*}(P_{yes})$, risk of symptom level for D_1)}. These should provide us with hints concerning presence of sure symptom levels and possible symptom levels in the case of D_1's recognition.

Let us utilize all theoretical assumptions in the next example to check the adjustment of medical data to the reasoning made above.

Example 3

We return to the primary information about patients belonging to P sampled in TABLE I to adopt only its part concerning data about P_1 and P_5. We design the table ($S_{B^2_{d_1}}$, $B^2_{d_1*}(P_{yes})$, risk of symptom level for D_1) as TABLE VII.

We decide the indiscernibility relation $I(B^2_{d_1*}(P_{yes})) = \{(S_8^{"l"}, S_8^{"l"}), (S_8^{"m"}, S_8^{"m"}), (S_8^{"h"}, S_8^{"h"}), (S_9^{"l"}, S_9^{"l"}), (S_9^{"m"}, S_9^{"m"}), (S_9^{"h"}, S_9^{"h"}), (S_8^{"l"}, S_8^{"h"}), (S_8^{"h"}, S_8^{"l"}), (S_8^{"l"}, S_9^{"l"}), (S_9^{"l"}, S_8^{"l"}), (S_8^{"l"}, S_9^{"h"}), (S_9^{"h"}, S_8^{"l"}), (S_8^{"m"}, S_9^{"m"}), (S_9^{"m"}, S_8^{"m"}), (S_8^{"h"}, S_9^{"l"}), (S_9^{"l"}, S_8^{"h"}), (S_8^{"h"}, S_9^{"h"}), (S_9^{"h"}, S_8^{"h"}), (S_9^{"h"}, S_9^{"l"})\}$. Its equivalence classes are listed in the following order: $IB^2_{d_1*}(P_{yes})(S_8^{"l"}) = IB^2_{d_1*}(P_{yes})(S_8^{"h"}) = IB^2_{d_1*}(P_{yes})(S_9^{"l"}) = IB^2_{d_1*}(P_{yes})(S_9^{"h"}) = \{S_8^{"l"}, S_8^{"h"}, S_9^{"l"}, S_9^{"h"}\}$ and $IB^2_{d_1*}(P_{yes})(S_8^{"m"}) = IB^2_{d_1*}(P_{yes})(S_9^{"m"}) = \{S_8^{"m"}, S_9^{"m"}\}$.

TABLE VII

($S_{B^2_{d_1}}$, $B^2_{d_1*}(P_{yes})$, RISK OF SYMPTOM LEVEL FOR D_1) IN SYMPTOM LEVEL CLASSIFICATION

Symptoms	Patients		Risk of Symptom Level for D_1
	P_1	P_5	
$S_8^{"l"}$	–	–	*no*
$S_8^{"m"}$	+	+	*yes*
$S_8^{"h"}$	–	–	*yes*
$S_9^{"l"}$	–	–	*no*
$S_9^{"m"}$	+	+	*yes*
$S_9^{"h"}$	–	–	*yes*

The set $(S_{B^2_{d_1}})_{yes} = \{S_8^{"m"}, S_8^{"h"}, S_9^{"m"}, S_9^{"h"}\}$ is surrounded by the lower approximation

$$(B^2_{d_1*}(P_{yes}))_*((S_{B^2_{d_1}})_{yes}) = \{S_8^{"m"}, S_9^{"m"}\}$$

and the upper approximation

$$(B^2_{d_1*}(P_{yes}))^*((S_{B^2_{d_1}})_{yes}) = \{S_8^{"l"}, S_8^{"m"}, S_8^{"h"}, S_9^{"l"}, S_9^{"m"}, S_9^{"h"}\}.$$

After interpreting the results we conclude that two levels $S_8^{"m"}$ and $S_9^{"m"}$ are sure markers of D_1 in a sample of patients waiting for their diagnoses when diagnosing according to the presence of S_8 and S_9. The classification provides us with possible levels including all members of $S_{B^2_{d_1}}$. The presence of low levels in the group of symptoms that may point out the existence of D_1 can be explained by the fact that code 1 has been placed both in the low and the medium levels of different symptoms, which has upgraded its power. High levels of symptoms are possible in D_1 but not very much representative since they rather warn against an attendance of more serious illnesses like myocardial infarct or stroke.

With respect to levels $S_8^{"m"}$, $S_9^{"m"}$ treated as truthful signals of D_1 and discovered in patients who also have been classified as truly suffering from D_1 in accordance to lower approximation sets, we should reconsider the last column in TABLE V. D_1 should be confirmed in these patients who represent the medium levels of both symptoms investigated; otherwise the decision about D_1's acceptance ought to be cautiously made as "*possible*". We thus re-edit TABLE V as TABLE VIII to state as sure classification of D_1 in the patients as possible by means of lower approximations of different rough sets involved.

TABLE VIII

(P, $B_{d_1}^2$, D_1) IN IMPROVED DIAGNOSIS CLASSIFICATION

Patients	S_8	S_9	Decision about D_1's presence
P_1	2	1	*yes* (*both levels are "medium"*)
P_2	3	2	*possible* (*risk for other severe diseases*)
P_3	1	1	*possible* (*low risk for* D_1)
P_4	3	2	*possible* (*risk for other severe diseases*)
P_5	2	2	*yes* (*both levels are "medium"*)
P_6	1	2	*possible*

We note that the remarks about the final decisions have become richer descriptions than formulations stated in the first version of TABLE VIII known as TABLE V. The expectations to meet D_1 in a patient are clearly expressed now, which means that the data, brought in by lower approximation sets, are the reliable sources of information.

If we wish to test other reducts or even the total set B in the same way then we should repeat the procedure described in Section IV to be furnished with a thorough prognosis concerning D_1.

V. CONCLUSIONS

The decision-makers, who have prepared a database concerning some clinical symptoms observed in a sample of patients, have been furnished with a mathematical apparatus known as rough sets.

To accomplish a classification of patients, due to a decision criterion accepted as the presence of a considered diagnosis, the technique engaging conditional and decision attributes has been used in order to verify and to correct the initial diagnostic hypotheses. The obtained classification lets us extract patients who surely suffer from the diagnosis as well as patients who surely or possibly are ill. These objects of the patient population are members of two approximation sets called the lower approximation and the upper approximation, respectively.

The classification has also allowed the decision-makers to exclude the category of patients who have been free from the illness in spite of heightened values of the observed symptoms. Even an undefined case of the patient has found its solution since the patient has been allocated in the group of possibly ill population members.

To avoid performing of too many numerical operations on the patients' reports concerning documented symptoms we have made a trial of introducing reduced groups of symptoms provided that the genuine results will be preserved. This has been accomplished by means of dependency rules referring to symptoms.

In the end, in order to add some complementary remarks to diagnostic decisions made by means of reducts we have proved the decisive character of symptom levels determined due to the symptom intensities. By testing a new classification accomplished on symptom levels with respect to patients belonging to the lower approximation set, we have found the levels that are safe indicators of the diagnosis occurrence in the patients. We can thus add complementary remarks to previously made diagnostic decisions in regard to the levels' presence or absence. The last stage of investigations has given us obvious hints concerning even prognoses of expecting another diagnosis in the patient.

All presented diagnostic models grounded on rough sets have been checked on small numbers of patients and symptoms to make the tests more understandable for a reader. In practice, we intend to check the designed methods by considering the large amount of data coming from the patients' reports that have been collected in one of the hospitals in the Blekinge district in Sweden.

REFERENCES

[1] J. Bazan, H. S. Nguyen and M. Szczuka, "A view on rough set concept approximations," *Fundamenta Informaticae*, vol. 59, 2004, pp. 107–118.

[2] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough based approaches," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, nr 12, 2004, pp. 1457–1471.

[3] T. Y. Lin and R. Chen, "Finding reducts in very large databases," *Proc. Joint Conf. Information Science Research*, 1997, pp. 350–362.

[4] S. K. Pal and A. Skowron (eds), *Rough Fuzzy Hybridization: New trends in Decision Making*, Singapore: Springer Verlag, 1999.

[5] S. K. Pal and P. Mitra, "Multi-layer perception, fuzzy sets and classification," *IEEE Trans. Neural Networks*, vol. 3, 1992, pp. 683–697.

[6] S. K. Pal and P. Mitra, "Case generation using rough sets with fuzzy representation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, nr 3, 2004, pp. 292–300.

[7] Z. Pawlak, "Rough sets," *Int. J. Computer and Information Science*, vol. 11, 1982, pp. 341–356.

[8] Z. Pawlak, "On rough sets," *Bulletin of the EATCS 24*, 1984, pp. 94–108.

[9] Z. Pawlak, *Rough Sets, Theoretical Aspects of Reasoning about Data*, Dordrecht: Kluwer Academic, 1991.

[10] Z. Pawlak, "Vagueness – a rough set view," *Structures in Logic and Computer Science*, 1997, pp. 106–117.

[11] Z. Pawlak, "Decision networks," *Proc. of Rough Sets and Current Trends in Computing*, Uppsala, Sweden, 2004, pp. 1–7.

[12] Z. Pawlak and A. Skowron, "Rough sets: some extensions," Information Sciences, vol. 177, 2007, pp. 26–40.

[13] E. Rakus, "Fuzzy set theory assisting medical diagnosis and appreciation of drug effectiveness," Doctor's dissertation, Medical Academy of Łódź, 1991 (in Polish).

[14] E. Rakus-Andersson, *Fuzzy and Rough Techniques in Medical Diagnosis and Medication*, Berlin Heidelberg: Springer Verlag, 2007.

[15] A. Skowron and C. Rauszer, "The discernibility matrices and functions in information systems, intelligent decision support," *Handbook of Applications and Advances of the Rough Set Theory* (A. Skowron ed.), Dordrecht: Kluwer Academic, 1992, pp. 331–362.

[16] J. T Yao and Y. Y.Yao, "Induction of classification rules by granular computing," *Proc. of the Third International Conference on Rough Sets and Current Trends in Computing (TSCTC'02)*, London, UK: Springer Verlag, 2002, pp. 331–338.

[17] Wang Changzhong and Hu Quinghua, "A new approach to attribute

reduction of consistent and inconsistent covering decision systems with covering rough sets," *Information Sciences,* vol. 177, issue 17, 2007, pp. 3500–3518.

[18] Xiangyang Wang, Jie Yang, Xiaolong Teng, Weijun Xia and R. Jensen, "Feature selection based on rough sets and particle swarm optimization," *Pattern Recognition Letters*, vol. 28, issue 4, 2007, pp. 459–471.

[19] Xizhao Wang, E. C. C. Tsang, Suyun Zhao, Degang Chen and D. S. Young, "Learning fuzzy rules from fuzzy samples based on rough set technique," *Information Sciences*, vol. 177, issue 20, 2007, pp. 4493-4514.