diagnostic principle, rule in database
data mining, optimal classification,
computer-based diagnostic system

Andrzej OSSOWSKI*

PRACTICAL ASPECTS OF COMPUTER-BASED CLINICAL DECISION SUPPORT SYSTEMS

The general problem of computer-based clinical decision support systems is considered. A class of rules in medical databases characteristic for therapeutic decisions has been distinguished. The necessity of application of a complete set of data for learning computerised systems of decision support has been pointed out. Rough set approach is applied to the analysis of the problem. An illustrative example of application of the presented results in the treatment of endometrial and breast cancer is given.

1. INTRODUCTION

Modern medicine and pharmacology offers to physicians a variety of drugs, methods and techniques that can be applied in clinical practice. Clinical treatment of a given disease is usually a complex, multistage therapeutic process with realisations dependent on particular cases. This is a typical situation for example in oncology [1]. At each stage of the process appropriate diagnostic patient's parameters are determined and physicians must make a decision concerning an optimal treatment option in the subsequent stage. The decision sequence associated with the therapy is not always exactly determined by medical standards but may be dependent also on certain random unpredictable factors and the knowledge and intuition of a given physician as well. Therefore the treatment details can be essentially different in different hospitals and the obtained result cannot be easily compared or applied to meta-analysis.

Computer-based systems of clinical decision support have been developed since many years [2,3]. Application of such systems, intended to help physicians in their clinical practice and standardise the decision process, is justified both by practical, medical and economical reasons. In many practical situations precise diagnostic decisions cannot be made on the basis of available measurable quantities. Then experts' knowledge, experience and intuition remain the best solution. In other situations rational diagnosis is possible but excluded by economical reasons if certain expensive procedures or procedures not available in a given hospital need to be applied. That is why

* Institute of Fundamental Technological Research, PAS, 00-049 Warszawa, ul. Świętokrzyska 21, aossow@ippt.gov.pl

it may be helpful especially for less experienced physicians to have access to an automatic system and confront their therapeutic decisions with those proposed by the system.

To realise a computer-based clinical decision support system it is necessary to collect certain diagnostic data and/or extract knowledge from data obtained from a group of experts [4]. Experts' knowledge is usually represented by a set of implicative IF-THEN rules while the diagnostic data contains records with proper therapeutic decisions. The computerised system is intended to minimise costs of expensive diagnostic procedures and standardise the decision process. Most papers on this subject are oriented towards the formal aspects of data mining and system learning. However, the problem of learning of computer-based decision support systems is not so simple in many practical applications and particularly in the case of therapeutic decisions. Exact learning data may not be available when optimal therapy of a given disease is not known in any particular case. This refers for example to anticancer chemotherapy with efficacy determined only statistically for a large group of patients [1]. Moreover, due to fundamental human psychological factors essential in the process of decision making also the knowledge extracted from a group of experts is incomplete in principle. There is no sense in learning a diagnostic classifier (e.g. a neural network or k-neighbours algorithm) by the use of an incomplete set of learning data. Any such a classifier even if well adjusted to the available data usually will not exhibit the property of generalisation i.e. its validation will not be successful, in general. That is why data completeness is crucial for designing of computer-based clinical decision support systems.

2. STRUCTURE OF CLINICAL DATABASES

It is common in clinical medicine that a therapeutic decision d, concerning a patient suffering from a given disease, is made on the basis of initial (independent) patient's attributes $x = (x_1,..., x_k)$ determined before a subsequent stage of the treatment. Initial attributes are usually assumed to belong to a bounded domain i.e. $x_i \in D_i$, i =1,...,k. In the result of the applied therapy T selected by the decision d certain final (dependent) patient's attributes $y = (y_1,..., y_l)$ are usually determined in order to estimate effects of the treatment and make subsequent decisions. Thus the database concerning a group of n patients is the following set of records

$$P = \{ p^i = [x^i , d^i, y^i], i= 1,...,n\} \qquad (1)$$

where each decision $d^i \in \{d_1,...,d_m\}$ (concerning the i-th patients) consists of choosing a therapy T^i from a finite set of treatment alternatives $T = \{ T_\alpha \}_{\alpha = 1,..., m}$. For any record $p^i \in P$ we denote by $x(p^i)$, $d(p^i)$, $y(p^i)$ the corresponding components x^i , d^i, y^i.

The described decision process can be illustrated by the following block scheme.

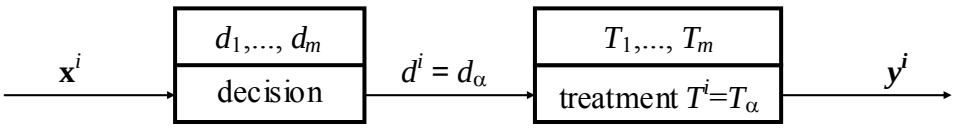

Fig.1. The block scheme of decision making in clinical practice.

Subsequent therapies are chosen on the basis of certain principles and rules justified by rational medical knowledge, standards and/or experts intuition.

3. RULES APPLIED IN CLINICAL PRACTICE

It is a typical situation that a physician (as an expert) asked to explain his diagnostic or therapeutic decisions becomes rather nervous and can provide only general, usually incomplete principles instead of precise rules. Nobody likes to accept the fact that his professional decisions could be explained by exact rules. On the other hand lack of such rules is a little upsetting and indicates that experts' decisions may be in part random i.e. not always unique and predictable. Thus decision processes in clinical practice may be in part intuitive and in part rational.

The question is what kind of decision rules are applied by physicians in clinical practice. It is logical to assume that the hidden diagnostic/therapeutic rules, if they exist, should be relatively simple. This hypothesis can be justified by fundamental features of human psychology. Indeed, if therapeutic decisions are generated consciously we may expect that the rules should have the standard form of IF -THEN rules where reasons are relatively simple logical sentences dependent on initial diagnostic parameters $x = (x_1,..., x_k)$. So, how does an expert make a therapeutic decision? Discussions with several physicians lead to the conclusion that experts do not look at diagnostic data simultaneously. Usually each diagnostic parameter is evaluated independently from the medical point of view and certain partial decisions are made and compared. Then a final therapeutic decision is made on the basis of the results of all partial decisions as is shown in Fig.1.

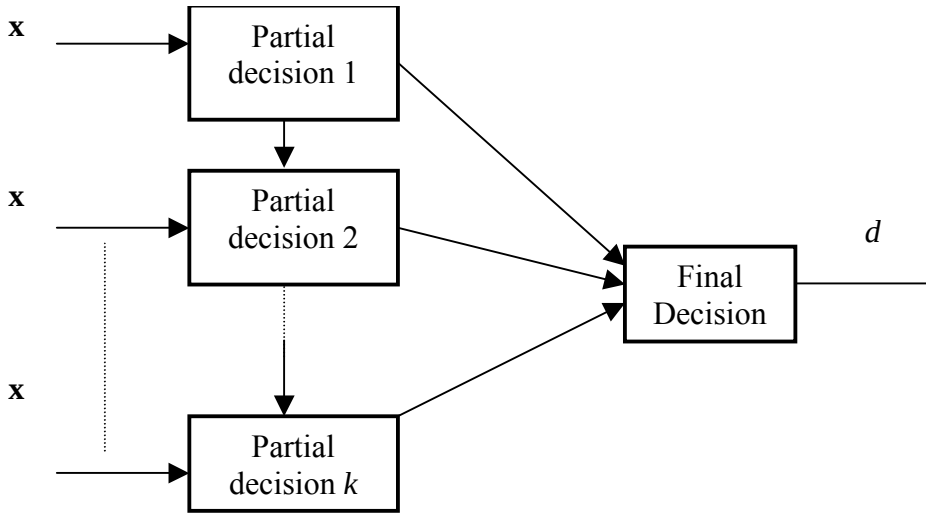

Fig.2. The block scheme of the decision process

Partial decisions are usually made in a certain order convenient for an expert or because of other medical factors. Each partial decision usually excludes certain therapeutic options and/or prefers other options. The first partial decisions reduce maximally the number of treatment options.

Since each partial decision is made on the basis of a single diagnostic parameter x_i we may conclude that the knowledge extracted from a group of experts and/or therapeutic data can be represented by a set of decision rules of the following form

$$R : IF\ (x_1 \in X_1) \wedge (x_2 \in X_2) \wedge \ldots \wedge (x_k \in X_k)\ THEN\ d \qquad (2)$$

where X_1 ,..., X_k are certain segments contained in the corresponding domains D_1 ,..., D_k. The domain of the practical rule R is a box $D(R) = X_1 \times \ldots \times X_k$. Each such a practical rule enables to accept a unique therapeutic decision d_α with a probability $p(R) \leq 1$ called rule certainty. More precisely, the rule certainty is the conditional probability that $d(p) = d$ for a randomly chosen record $p \in P$ satisfying the condition $x(p) \in X_1 \times \ldots \times X_k$. It is also possible that different rules of form (2) enable to choose the same treatment option T_α . Then the corresponding decision region is approximated by a sum of boxes contained in the space $D_1 \times \ldots \times D_k$.

One can pose another question concerning the total number of practical rules used by a physician i.e. how many practical rules can be extracted from clinical data? It is well known that human brain is able to perceive simultaneously about seven objects. For this reason the total number of therapeutic options should be small and we may expect a few (no more than seven) main rules each dependent on a few parameters (attributes). The remaining rules discovered in a given database are probably fictitious and have no practical sense. Thus there is a problem of selection of the proper rules from the collection of all rules discovered in a given database. This problem can be analysed by using a concept of data incompleteness which is considered in the next section.

4. INCOMPLETENESS OF LEARNING DATA

Patients suffering from a given disease can be divided into groups according to different treatment options applied in a particular stage of the therapeutic process. Making proper decisions on the treatment options is thus a kind of classification of objects determined by a finite set of attributes $x = (x_1,\ldots,x_k)$. A computer-based system for clinical decision support can apply directly the discovered practical rules or can be learned on the basis of the available therapeutic data P. It is usually believed that a large number of records in the learning database provided by experts is sufficient to build up a good classifier for clinical decision support. Unfortunately this is not true in general. Independently on the number of records learning data should be complete (representative) in this sense that the learnt classificator is able to generalise the stored data. This is possible when the domains of all practical rules cover the whole space $D_1 \times \ldots \times D_k$. Then all cases can be diagnosed and appropriate treatment option can be chosen. Unfortunately, a collection of practical rules discovered in a given medical database may not be complete and rules may not be exact. This means that the same or very similar cases may be diagnosed and treated in different ways what cannot be easily accepted even by patients. The probability of such situations should be minimised. However, a random factor cannot be excluded from the diagnostic process. It is also possible that patients with similar initial attributes treated by the same therapy T_α tend to quite different final attributes after the treatment. Such a situation may be an indication that the applied collection of initial attributes is not sufficient to describe the therapeutic process of the disease. The generalisation abilities of a artificial decision support system will be limited.

How to measure data (in)completeness? To describe completeness of clinical database P let us assume that certain rules $R_1,\ldots,R_m$ of form (2) have been extracted and selected as those having practical (clinical) meaning. Then the most natural measure of data completeness is the following ratio

$$c(P) = card\ (D(R_1) \cup \ldots \cup D(R_m))/\ card\ (D_1 \times \ldots \times D_k) \qquad (3)$$

i.e. percentage of cases covered by the selected rules $R_1,...,R_m$. Thus the factor c(P) determines the rules strength. Obviously, data incompleteness can be expressed by the factor 1- c(P). This is a separate problem which of the extracted rules should be selected and what levels of data incompleteness can be accepted. The selection can be performed with respect to rules certainty and certain medical suggestions. The lower bound on the minimal acceptable completeness ratio c(P) can be deduced from simple geometric conditions as equal to the maximal relative volume of a m-dimensional box contained in a fixed m-dimensional sphere.

The introduced value c(P) characterising a given clinical database is not always sufficient to describe data completeness. Not only rule strength but also rules certainty should be taken into account in order to determine randomness of therapeutic decisions. It is clear that practical rules of form (2) approximate any real decision set by boxes in the decision space. The approximation accuracy is dependent on an expert experience/intuition. A lower and upper approximation can be distinguished to determine sets of exact and random decisions, respectively. Domains of different rules can intersect. The above conjectures can be easily described in terms of the rough set theory.

5. ROUGH SETS APPROACH

In clinical practice usually discrete patient's attributes are used. Some attributes are discrete in principle for example patient sex or the information whether a woman suffering from breast cancer intended for pharmacological treatment is or is not after mastectomy. Other attributes which are continuous in their nature (e.g. patient's age, hormonal state, clinical state) are usually discretised for convenience. Attributes discretisation simplifies the treatment description, clinical standards and decision making process. However, too many levels of discretisation may cause the decision rules to become very complex and impractical. On the other hand, the decision rules may be too imprecise if the assumed resolution of a discrete scale is too small. Therefore, levels of a discrete scale for a given attribute should be chosen in some sense optimally to coincide with certain medical evidence, clinical clearness and practical convenience. In particular, a discrete scale may be dependent on the disease (For example three or four age periods are applied for breast cancer patients to distinguish women in different hormonal states [1]).

The rough sets theory is especially adequate to describe the problem of rules and data incompletness in databases with discrete attributes [7, 8, 9]. Indeed, let us denote by P_α the set of all records $p \in P$ for which the attribute $d = d_\alpha$ i.e.

$$P_\alpha = \{ p = (x, d, y) \in P : d = d_\alpha \} \qquad (4)$$

and corresponds to the set of patients who are treated according to the therapy T_α . It is clear that $P_\alpha \cap P_\beta = \phi$ and $P = P_0 \cup \ldots \cup P_m$. Thus the attribute d determines a partition Π (P, d) of the total space P into non-intersecting classes. We denote by Π (p, d) the class containing the record $p \in P$. It is interesting to find out to how extend classes Π (p, d), $p \in P$ are determined by initial decision attributes x. Let $X \subset P$ be a set of records $p = (x, d, y) \in P$ satisfying certain logical condition $\Phi(x)$ i.e.

$$X = \{ p = (x, d, y) \in P : \Phi(x) \} \quad (5)$$

The rough set theory is based on the concepts of Π - lower and Π - upper approximation (i.e. with respect to a partition Π) of any set X⊂P which are defined as follows:

$$\Pi_{\#}(X) = \{p \in P : \Pi(p, d) \subseteq X\}, \Pi^{\#}(X) = \{p \in P : \Pi(p, d) \cap X \neq \phi\}.$$

The difference $B_{\Pi}(X) = \Pi^{\#}(X) - \Pi_{\#}(X)$ is called the Π - boundary of the set X . If $B_{\Pi}(X) = \phi$ then the set X is said to be crisp (exact) with respect to the partition D. In practice this means that the logical condition Φ(x) dependent on initial attributes enables us to make a unique therapeutic decision d. In other words, all patients with the same initial attributes are treated in the same way. Unfortunately in clinical practice such a situation is not the case in general. Certain diagnostic attributes x take continuous values and the boundary between the choice of one or another therapy is not sharp. That is why the decision sets in the total space P are usually rough sets.

To describe such imprecise decision rules one can apply the rough membership function defined as follows:

$$\mu_{D,X}(p) = card(\Pi(p, d) \cap X) / card(X). \quad (6)$$

The membership function determines the probability that a given record p ∈ X belongs to the class Π (p, d). Similarly, we can define the probability of false (correct) classification of a record.

It is easy to find out that the quantity

$$\nu_{D,X}(p) = card(\Pi(p, d) \cap X) / card(\Pi(p, d)). \quad (7)$$

is simply the probability that the record p will be classified as X employing the set of attributes x.

It is clear that any rough approximation of a decision set corresponds to a set of decision rules of the form IF-THEN. All essential rules characteristics such as rule strength, coverage, certainty can be expressed in terms of rough set theory. Rough analysis of a clinical database enables to determine data incompleteness and distinguish the most questionable records (with relatively small values of membership functions) belonging to rough boundaries. The higher is the roughness of decision sets the higher is data incompleteness.

The rough set theory also enables to analyse the so-called dependence of attributes. It is possible in clinical databases that certain group of attributes is dependent on another group of attributes. Such a dependence is rarely crisp i.e. there is no exact functional relation between the two groups. The more realistic is a rough dependence of attributes which can be defined by a rough measure. On this basis it is possible to perform a reduction of attributes which is sometimes possible and/or desirable in order to simplify and improve decision-making [5, 6].

6. OPTIMALITY OF THE SET OF THERAPEUTIC RULES

The output attributes y are obviously dependent on initial attributes x and the applied therapy T_{α} . Looking for the dependence/relations between (x, d) and y is a separate medical problem. In

diagnostic systems the most important is the clinical effectiveness of the applied decision rules and therapies T_α , $\alpha = 0,1,..,k$. We call the decision rules and therapies the treatment program.

The therapeutic effectiveness Q of a treatment program applied to a given patient is a function of the initial and final patient's attributes i.e. Q=Q(x, y). The therapeutic effectiveness averaged over a large group of patients is of clinical importance. The averaged value $< Q >$ is dependent on the set of rules applied by experts. The rules are usually not only incomplete but also sub-optimal in the sense of averaged final effects. In many cases essential final patient's attributes are not registered and referred to undertake therapeutic decisions. This is typical in the case of long term treatment and effects (e.g. five years survival in oncology.). The treatment optimisation should be an adaptive iterative process inducing appropriate changes of rules parameters that ensures minimisation of the assumed quality(therapeutic effectiveness) index of the treatment.

In modern medicine also economical factors are essential. Then a pharmaco-economical optimisation index dependent on the applied therapy Q = Q(x, d, y) should be optimised. Such an approach is justified especially in the case when different treatment programs yield very similar therapeutic effectiveness. Then one can accept application of the therapeutic program that ensures minimal cost of the treatment. This is of course dependent on all decisions d^i and total costs of the applied therapies T_α.

7. COMPUTER-BASED CLINICAL DECISION SUPPORT SYSTEMS

In this section we provide the general methodological background of computer-based diagnostic/therapeutic decision support systems that summarises our previous considerations.

The first step in designing of a good computer-based diagnostic/therapeutic decision support system is extraction of knowledge and therapeutic data from a group of experts. If the experts cannot formulate exact decision rules appropriate data mining algorithms should be applied to find rules in the available database. The general principles described in Section 3 enable to limit the class of rules extracted from the data. A few rules of relatively high certainty and coverage should be selected as practical rules having a clinical meaning. The remaining rules should be treated as accidental ones. Once the practical rules are selected their completeness and uncertainty should be determined. If a very small percentage of cases cannot be explained by the rules the available data can be applied in learning of a computer-based classificator. However, if a lot of cases cannot be correctly diagnosed then one should assume that the collected data are not complete.

Rough set analysis of data should be performed in order to distinguish rough boundaries of decision sets. Questionable records should be discussed with the experts in order to interpret data incompleteness. It is valuable to ask the experts about lower and upper approximations i.e. what are the critical values of attributes determining unique decisions and what are the values that exclude certain decisions. New attributes should be included into the databases if necessary. Then by using the rough set approach the analysis of attributes dependence should be performed and the possibility of attribute reduction should be considered.

Finally, an appropriate software should be applied to learn a diagnostic computer classificator based for example on neural networks or the method of k - neighbours [10]. It can be expected that a computer-based classificator enables to achieve better results than a system applying directly a set of implicative decision rules.

The computer-based system designed according to the above methodology is usually suboptimal. Long-term effects of the therapeutic program should be registered and applied to further system optimisation. Both parameters of decision rules and levels of the applied discrete scales should be adjusted in order to improve an assumed quality index in each optimisation step.

8. APPLICATION TO ANTICANCER THERAPY

Many cancers allow a lot of different treatment options so that making therapeutic decision is an essential element of anticancer therapies. In particular, women with diagnosed operable breast or endometrial cancer are usually intended to chemotherapy, hormonal therapy or palliative therapy but it is also possible that no further standard treatment after surgery is applied. Everything is dependent on certain factors and the general clinical stage of a given patient, in particular. Basic discrete parameters such as tumor grade I,II,III,IV, clinical state of the patient and discrete TNM classification of cancer advancement are helpful for physicians to make a proper decision. There are also applied at least two or three age segments which are biologically justified (for example women before and after menopause).

A small clinical database containing about 200 records concerning women suffering from endometrial cancer was analysed by the author. Data mining enables to find a number of rules in the base. The selected group of 5 rules with relatively large certainty proved to be incomplete. Therapeutic decisions of about 40% of records cannot be explained by the rules. Further consultations show that certain essential details, obvious for physicians but not for computers, were not included into the database. This was the main reason of data incompleteness. Additional attributes included into the database enable to improve data completeness so that the idea of a computer-based diagnostic system becomes more realistic.

An appropriate discretisation of patients' age had proved to be essential. Two and three optimal age levels have been proposed for patients suffering from endometrial cancer. Further studies are necessary to answer the question whether the presence of estrogen and/or progesterone receptors in cancerous cells should be taken into account in the decision process similarly as in the case of breast cancer.

9. CONCLUSIONS

Various practical aspects concerning computer-based clinical decision support systems have been discussed. A crucial role of data (in)completeness in learning has been pointed out. It has been shown that there is no practical sense to learn an artificial classifier by using incomplete clinical data. The existence of practical rules applied by professionals has been postulated on the basis of fundamental psychological considerations. Data mining techniques oriented towards discoveries of practical rules in medical databases enables us to determine data (in)completeness before learning of the system.

BIBLIGRAPHY

[1] MADEJ G., Chemioterapia onkologiczna dorosłych i dzieci, PZWL, Warszawa 1994.
[2] REISMAN Y. Computer-based clinical decision aids. A review of methods and assessment of systems. Med Inf (Lond) 21(3):179-97, Jul-Sep, 1996.
[3] SHORTLIFFE EH, DAVIS R, AXLINE SG, BUCHANAN BG, GREEN CC, COHEN SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res 1975 Aug;8(4):303-20
[4] WOŹNIAK M., PUCHAŁA E., Zastosowanie metody generowania reguł eksperta w zadaniu diagnostyki medycznej, X Krajowa Konferencja Naukowa, Biocybernetyka i Inżynieria Biomedyczna, tom II, 679-683, Warszawa, 1977.
[5] JANECKI J., Znaczenie optymalizacji doboru cech w wielowymiarowych systemach diagnostycznych, X Krajowa Konferencja Naukowa, Biocybernetyka i Inżynieria Biomedyczna, tom II, 656-662, Warszawa, 1977.
[6] JANECKI J., Computer selection of differential diagnostic parameters, "Information systems for patient care", J. van Egmont, North Holland Publ. Corp., 401-404, 1976.
[7] PAWLAK Z., Rough sets, International Journal of Computer Information Sciences, 11, 341-356, 1982.
[8] PAWLAK Z., Rough classification, International Journal of Man-Machine Studies, 20, 469-483, 1984.
[9] PODRAZA W., Modyfikacja zastosowania teorii zbiorów przybliżonych dla celów medycznych, X Krajowa Konferencja Naukowa, Biocybernetyka i Inżynieria Biomedyczna, tom II, 532-536, Warszawa, 1977.
[10] KURZYŃSKI M., SAS J., WIKIERA I., Rule-Based Medical Decision-Making with Learning, Proc. 12th World IFAC Congress, vol. 4, 334-338, Sydney, 1993.