Integrated consistent and complete "expert" and data mining

Boris Kovalerchuk[1], Evgenii Vityaev[2]. James F. Ruiz[3]

[1]Department of Computer Science, Central Washington University, Ellensburg, WA, 98926-7520, USA, borisk@cwu.edu

[2]Institute of Mathematics, Russian Academy of Sciences, Novosibirsk 630090, Russia vityaev@math.nsc.ru

[3]Department of Radiology, Woman's Hospital, Baton Rouge, LA 70895-9009, USA, MDJR@womans.com

1. INTRODUCTION.

Integration of knowledge management (KM) and Data Mining (DM) methods can benefit many applications. It permits to combine and mutually verify knowledge obtained from experts and extracted from raw data. Traditional expert systems rely on knowledge "extracted" in the form of If-Then diagnostic rules from experts. Systems based on Machine Learning technique rely on an available database for discovering diagnostic rules. These two sets of rules may contradict each other. An expert may not trust rules, as they may contradict his/her existing rules and experience. Also, an expert may have questionable or incorrect rules while the data/image base may have questionable or incorrect records. Moreover, data mining discovery may take the form different from If-Then rules and these rules may need to be decoded before they are compared to expert rules.

For high-risk applications such as financial investment and life-critical medical applications, e.g., for breast cancer diagnosis, benefits of wise integration are especially evident [Kovalerchuk, Vityaev, Ruiz, 2000]. Suppose that a DM method extracted an "excellent" diagnostic rule, which is near 100% correct on the data used for discovering and testing the rule. On the other hand, assume that this rule contradicts the opinion of an experienced expert. Who will risk relying on such rule without extra analysis, e.g., for cancer diagnosis? An expert can argue that the rule was extracted from a non-representative database (DB) even if it is a large one. For instance, the DB contains a huge amount of negative examples (mammograms of benign cases) and just few positive (cancer) cases. Not the full variety of positive cases may be fully represented in the database. Similarly in risky investment, a buy/sell signal generated by a DM method may contradict an opinion of an experienced trader/investor. Integration of DM and KM methods can help to identify such situations beforehand, saving lives, money and bringing other benefits. Next, working on producing a consistent result may reveal a source of contradiction (e.g., a non-representative database or an unmotivated expert's opinion). Finally, this will build a foundation for better-combined results.

In this paper knowledge management and data mining techniques are integrated using two methods -- one from the KM area and other one from the DM area. The methods and their combination are powerful and unique in some sense. The first method is focused on knowledge acquisition from humans directly via dynamic optimized expert interview. It is called Boolean "Expert" Mining (BEM- [Kovalerchuk, Vityaev, 2000, ch. 3]. The second method called MMDR (Machine Method for Discovering Regularities [Kovalerchuk, Vityaev, 2000, ch. 4]) extracts knowledge from raw data using relational approach. The power and uniqueness of these methods and their combination is coming from three their properties -- methods are (1) complimentary, (2) consistent and (3) complete.

The methods **compliment** each other because BEM leverages human knowledge of a problem and MMDR leverages patterns hidden in raw data. The integrated BEM and MMDR methods produce **consistent knowledge**, i.e., knowledge free of contradictions (between rules generated by each of them separately and together. Specifically in a medical application discussed below data mining results are consistent with rules used by an experienced medical expert and a database of pathologically confirmed cases. Similarly **complete methods** produce complete knowledge systems (models), i.e., models which classify all (or largest possible number of) combinations of the used attributes.

Integration of data mining and knowledge management methods requires resolving many still open issues at the junction of the two fields. Below we discuss them and outline our approach.

The representational mismatch. "Expert" mining method extract knowledge from expert, in contrast data mining methods that discover knowledge from data. Often it is difficult to compare such knowledge, because of different representations. For instance how to compare and conclude consistency/inconsistency of a Neural Network and expert IF-THEN rules? To solve this problem we use a relational rule-based approach in the data mining part of integration [Kovalerchuk, Vityaev, 2000]. Thus, DM and "expert" mining both produce rules. These rules are interpretable by humans, which is not so obvious for Neural Networks and Discriminant Analysis. This relational approach permits interpretable and readable representation of any data types and hypothesis. The representative measurement theory [Krantz, et al. 1971, 1989, 1990] is used as a tool for interpretable relational representation of various data types in the first-order logic.

Guiding the knowledge discovery process. This process must ensure that the result is a consistent non-contradictory rule base that includes both expert rules and rules extracted using DM. Our approach employs monotonicity in expert mining to avoid contradictions. Similar idea used in the data mining algorithm MMDR starting from simplest logical expressions and adding more clauses.

Incremental learning and knowledge assimilation. Such process is an important component of integration. It allows to ensure that with more data and more interviews of experts consistent knowledge will increase and finally reach complete knowledge for a given language. In this context completeness meant that for every case described in the language the system has a rule for classifying the case. It is proved for both methods that they can find complete sets of rules for both an expert and a representative data set.

Dealing with the qualitative of knowledge. Much (if not majority) of expert knowledge is qualitative knowledge. However, it is not so common for knowledge discovered by data mining methods. Often this knowledge is quantitative and should be interpreted by experts for comparing and integration with qualitative knowledge. The representative measurement theory shows the way in which quantitative knowledge can be converted into qualitative knowledge without loss of information [Krantz et al, 1971, 1989, 1990; Kovalerchuk, Vityaev, 2000]. This conversion is based on the idea that qualitative knowledge is a numerical representation of some relational structures. In [Kovalerchuk, Vityaev, 2000] we show how ordinary representations of data types such as comparisons, binary matrices, matrices of orderings, matrices of proximity and attribute-based matrix can be described in relational form without loss any information. It requires an appropriate relational language for representing data and hypotheses tested.

These problems make the design of an integrated DM/KM system extremely complex and raises two additional complex tasks:

(1) Identify contradictions between expert diagnostic rules and knowledge discovered by data mining mechanisms and

(2) Eliminate contradictions between expert rules and machine discovered rules.

If the first task is solved, the second task can be approached by cleaning the records in the database, adding more features, using more sophisticated rule extraction methods and testing the competence of an expert. If rule extraction is performed without these tasks in mind, it is difficult to recognize a contradiction. In addition, rules generated by an expert and data-driven rules maybe incomplete as they may cover only a small fraction of possible feature combinations. This can make impossible to confirm that rules are consistent within an available database. Additional new cases or features can make the contradiction apparent. Therefore, the major problem here is *discovering sufficient, complete, and comparable sets of expert rules and data-driven rules*. Such completeness is critical for comparison. For example, suppose that expert and data-driven rules cover only 3% of possible feature combinations (cases). If there are no contradictions between these rules, there is still plenty of room for contradiction in the remaining 97% of the cases.

Discovering complete set of regularities/rules. If data mining method X discovers an incomplete set of rules, R(X), then rules R(X) do not produce an output (forecast) for some inputs. If two data mining methods, X and Y, produce incomplete sets of rules R(X) and R(Y) then it would be difficult or impossible to compare them if R(X) has a rule for input **a,** but R(Y) does not. Similarly, an expert mining method, E, can produce a set of rules R(E) with very few rules overlapped in R(X) and R(Y). Again, this creates a problem in comparing the performances of R(X), R(Y) and R(E). Therefore, completeness is a very valuable property for any data mining method. The Boolean Expert Mining method (BEM) described below is aimed to build a complete set of rules in a given language.

If an expert has a judgment about a particular type of patient symptom or symptom complex then appropriate rules can be extracted by BEM.

The problem is to find a method, W, such that R(W)=R(X)UR(Y) for any X and Y, i.e., this method, W, will be the most general for a given data set. The MMDR is a complete method for relational data in this sense. If data are not sufficient, then MMDR utilizes the available data and attempts to keep statistical significance within an appropriate range. This method attempts to maximize the domain of the rules. In other words, the BEM method extracts a complete set of rules from a expert and the data mining method MMDR extracts a complete set of rules from data. For MMDR and BEM, this has been proved in [Vityaev, 1992; Kovalerchuk et al. 1996].

Thus, the first goal of this paper is to present methods for discovering complete sets of expert rules and data-driven rules. Unfortunately, this objective leads us to an exponential, non-tractable problem of extracting diagnostic rules. A brute-force method may require asking the expert thousands of questions. For example, for 11 binary diagnostic features of clustered calcifications found in mammograms, there are (2^{11}=2,048) feature combinations, each representing a unique case. A brute-force method would require questioning a radiologist on each of these 2,048 combinations. A related problem is that experts may find it difficult or impossible to articulate confidently the large number of interactions between features. Dhar and Stein [1997] pointed out that If a problem is "*decomposable*" (the interactions among variables are limited) and experts can articulate their decision process, a rule-based approach may scale well. An effective mechanism for decomposition-based monotonicity is presented below.

Creating a **consistent integrated rule base** includes the following **steps:**

1. Finding data-driven rules not discovered by asking an expert.
2. Analysis of these new rules by an expert using available proven cases. A list of these cases from the database can be presented to an expert. The expert can check:

2.1 Is a new rule discovered because of misleading cases? The rule may be rejected and training data can be extended.

2.2. Does the rule confirm existing expert knowledge? Perhaps the rule is not sufficiently transparent for the expert. The expert may find that the rule is consistent with extert's previous experience, but the expert would like more evidence. The rule can increase the confidence of expert's practice.

2.3. Does the rule identify new relationships, which were not previously known to the expert? The expert can find that the rule is promising.

3. Finding rules which are contradictory to the experts knowledge or understanding. There are two possibilities:

3.1. The rule was discovered using misleading cases. The rule must be rejected and training data must be extended.

3.2. The expert can admit that his/her ideas have no real ground. The system improves expert experience.

2. METHOD FOR DISCOVERING DIAGNOSTIC RULES FROM DATA BASE

The MMDR method expresses patterns in first order logic and assigns probabilities to rules generated by composing patterns. Learning systems based on first-order representations have been successfully applied to many problems in chemistry, physics, medicine, finance and other fields [Mitchell, 1997, Russell, Norvig, 1995]. As any technique based on logic rules, this technique allows one to obtain human-readable forecasting rules that are **interpretable** in a particular field. For instance a medical expert can evaluate the correctness of the diagnosis as well as a diagnostic rule. The critical issue in applying data-driven forecasting systems is generalization. MMDR and related "Discovery" software systems [Vityaev, Moskvitin, 1993] generalize data through "law-like" logical probabilistic rules.

Conceptually, **law-like rules** came from philosophy of science. These rules attempt to mathematically capture the essential features of scientific laws**:** (1) high level of generalization; (2) simplicity (Occam's razor); and, (3) refutability. The first feature -- generalization -- means that any other regularity covering the same events would be less general, i.e., applicable only to a subset of events covered by the law-like regularity. The second feature – simplicity--reflects the fact that a law-like rule is shorter than other rules. The law-like rule (R1) is more refutable than another rule (R2) if there are more testing examples which refute (R1) than (R2), but the examples fail to refute (R1).

Formally, we present an IF-THEN rule C as $A_1\& \ldots \& A_k \Rightarrow A_0$, where the IF part, $A_1\&...\&A_k$, consists of true/false logical statements $A_1,\ldots,A_k$,and the THEN part consists of a single logical statement A_0. Statements A_i are some given refutable statements or their negations, which are also refutable. Rule C allows us to generate sub-rules with a truncated IF part, e.g. $A_1\&A_2 \Rightarrow A_0$, $A_1\&A_2\&A_3 \Rightarrow A_0$ and so on. For rule C its conditional probability Prob(C) = Prob($A_0/A_1\&...\&A_k$) is defined. Similarly conditional probabilities Prob($A_0/ A_{i1}\&...\&A_{ih}$) are defined for sub-rules Ci of the form $A_{i1}\& \ldots \& A_{ih} \Rightarrow A_0$.

Conditional probability Prob(C) = Prob($A_0/A_1\&...\&A_k$) is used for estimating forecasting power of the rule to predict A_0. The **rule is** "**law-like**" iff all of its sub-rules have a statistically significant **lower conditional probability** than the rule. Each sub-rule C_i generalizes rule C, i.e., potentially C_i is true for a larger set of instances. Another definition of "law-like" rules can be stated in terms of generalization. The rule is "law-like" iff it can not be generalized without producing a statistically significant reduction in its conditional probability**.** "Law-like" rules defined in this way hold all three properties of scientific laws. They are: (1) general from a logical perspective, (2) simple, and (3) refutable. Section 4 presents some rules extracted using this approach.

MMDR searches all chains C_1 , C_2 , …, C_{m-1}, C_m of nested "law-like" subrules, where C_1 is a subrule of rule C_2 , C_1 = sub(C_2), C_2 is a subrule of rule C_3, C_2 = sub(C_3) and finally C_{m-1} is a subrule of rule C_m, C_{m-1} = sub(C_m). Also Prob(C_1) < Prob(C_2), … , Prob(C_{m-1}) < Prob(C_m). There is a **theorem** [Vityaev, 1992] that **all rules, which have a maximum value of conditional probability, can be found at the end of such chains**. The algorithm stops generating new rules when they become too complex (i.e., statistically insignificant for the data) even if the rules are highly accurate on training data. The Fisher statistical criterion is used in this algorithm for testing statistical significance. The obvious other stop criterion is time limitation. Theoretical advantages of MMDR generalization are presented in [Kovalerchuk, Vityaev, 2000, Vityaev, 1992]. Note that a class of general propositional and first-order logic rules, covered by MMDR is wider than a class of decision trees [Mitchell, 1997].

3. METHOD FOR EXTRACTING DIAGNOSTIC RULES FROM EXPERT

3.1. Hierarchical Approach

The interview of an expert to extract rules is managed using an original method of monotone Boolean function restoration [Hansel, 1966, Kovalerchuk et al, 1996]. One can ask an expert to evaluate a particular case when a number of features take on a set of specific values. A typical query will have the following format:

- "If feature 1 has value V_1, feature 2 has value V_2, ..., feature n has value V_n, then should biopsy be recommended or not?
- Or, does the above setting of values correspond to a case suspicious of cancer or not?"

Each set of values (V_1, V_2, ...,V_n) represent a possible clinical case. It is practically impossible to ask an expert radiologist to generate diagnosis for thousands of possible cases. **A hierarchical approach** combined with the use of the property of **monotonicity** makes the problem manageable.

We construct a hierarchy of medically interpretable features from a very generalized level to a less generalized level. This hierarchy follows from the definition of the 11 medically oriented binary attributes. The medical expert indicated that the original 11 binary attributes w_1, w_2, w_3, y_1, y_2, y_3, y_4, y_5, x_3, x_4, x_5 could be organized in terms of a hierarchy with development of two new generalized attributes x_1 and x_2:

Level 1 (5 attributes)		Level 2 (all 11 attributes)
x_1 ("Amount and volume of calcifications")	–	w_1, w_2, w_3
x_2 ("Shape and density of calcification")	–	y_1, y_2, y_3, y_4, y_5
x_3 ("ductal orientation")	–	x_3
x_4 ("comparison w. previous examination")	–	x_4
x_5 ("associated findings")	–	x_5,

Five binary features x_1, x_2, x_3, x_4, and x_5, constitute level 1. A new generalized feature, with grades (0 - "benign" and 1 - "cancer") was introduced based on features: w_1—number of calcifications/cm^3; w_2--volume of calcification, cm^3; w_3--total number of calcifications. Variable x_1 is viewed as a function $\nu(w_1, w_2, w_3)$ to be identified. Similarly a new feature with grades: (1) for "marked" and (0) for

"minimal" or, equivalently (1)-"cancer" and (0)-"benign" generalizes features: y_1 -- "Irregularity in shape of individual calcifications"; y_2 -- "Variation in shape of calcifications"; y_3 -- "Variation in size of calcifications"; y_4 -- "Variation in density of calcifications"; y_5 -- "Density of calcifications". Variable x_2 is viewed as a function $x_2=\psi(y_1, y_2, y_3, y_4, y_5)$ to be identified for cancer diagnosis. A similar structure was produced for a decision regarding biopsy. The expert was requested to review both the structure and answers for the questions:

"Can function f_1 be assumed the same for both problems? ,

"Can function f_2 be assumed the same for both problems? ,

The expert indicated that these two functions ν and ψ should be common to both problems:

(P1) recommendation biopsy and (P2) cancer diagnosis. Therefore, the following relation is true regarding the f_i (for i = 1, 2) and the two φ, and ψ functions:

$$f_i(x_1,x_2,x_3,x_4,x_5) = f_i(\varphi(w_1,w_2,w_3), \psi(y_1,y_2,y_3,y_4,y_5), x_3,x_4,x_5), \; i = 1,2.$$

Further levels of hierarchy can be developed for better describing the problem. For example, y_1 ("irregularity in shape of individual calcifications") may be found in 3 grades: "mild" (or t_1), "moderate" (or t_2) and "marked" (or t_3). Next observe that it is possible to change (i.e., generalize) the operations used in the function $\psi(y_1,y_2,..,y_5)$. For instance, we may have mentioned function ψ as follows: $\psi(y_1,y_2,..,y_5) = y_1 \,\&\, y_2 \vee y_3 \,\&\, y_4 \,\&\, y_5$, where & and $\vee$ are the binary, logical operations for "AND" and "OR", respectively. Then, & and $\vee$ can be substituted for one of their multivalued logic analogs, for example, $x \,\&\, y = \min(x,y)$ and $x \vee y = \max(x,y)$ as in fuzzy logic. We assume that x_1 is the number and the volume occupied by calcifications, in a binary setting, as follows: (0-"against cancer", 1-"for cancer"). Similarly, let: x_2,x_3,x_4,x_5--with values: 0-"benign", 1-"cancer".

3.2. Monotonicity

To understand how a mammogram. Given the above definitions we can represent clinical cases in terms of binary monotonicity is applied we use the same breast cancer problem -- the evaluation of calcifications in vectors with five generalized features as: (x_1,x_2,x_3,x_4,x_5). Next consider the two clinical cases that are represented by the two binary sequences: (10110) and (10100). If one is given that a radiologist correctly diagnosed (10100) as a malignancy, then, by utilizing the property of monotonicity, we can also conclude that the clinical case (10110) should also be malignancy.

This conclusion is based on the **systematic coding of all features** "suggestive for cancer" as **1**. Observe that in (10100) we had two indications for cancer:

$x_3=1$ (ductal orientation having value of 1; suggesting cancer) and

$x_1=1$ (Amount and volume of calcifications with value 1 indicating cancer).

In the second clinical case we have these two observations for cancer and also $x_4=1$ (a comparison with previous examinations suggesting cancer). In the same manner if we know that (01010) is not considered suspicious for cancer, then the case (00000) should also not be considered suspicious. This is true because in the second case we have less evidence indicating the presence of cancer. The above considerations are the essence of how our algorithms function. They can combine logical analysis of data with monotonicity and generalize accordingly. In this way, the weaknesses of the brute-force methods can be avoided. It is assumed that if the radiologist believes that the case is malignant, then he/she will recommend a biopsy. More formally, these two sub-problems are defined as follows:The Clinical Management Sub-Problem (P1): One and only one of the following two disjoint outcomes is possible:1) "Biopsy is necessary", or: 2) "Biopsy is not necessary".

The Diagnosis Sub-Problem (P2): Similarly as above, one and only one of two following two disjoint outcomes is possible. That is, a given case is: 1) "Suspicious for malignancy", or:2) "Not suspicious for malignancy". Our goal here is to extract the way the system operates in the form of two discriminant Boolean functions f_2 and f_1:

Function f_1 returns true (1) value if the decision is "biopsy is necessary", false (0) otherwise.

Function f_2 returns true (1) value if the decision is "suspicious for malignancy", false (0) otherwise.

The first function is related to the first sub-problem, while the second function is related to the second sub-problem. There is an important relation between these two sub-problems P1 and P2 and functions $f_1(\alpha)$, $f_2(\alpha)$. The problems are nested**,** i.e., if the case is suggestive of cancer ($f_2(\alpha)=1$) then biopsy should be recommended ($f_1(\alpha)=1$) for this case, therefore $f_2(\alpha)=1 \Rightarrow f_1(\alpha)=1$. Also if biopsy is not recommended ($f_1(\alpha)=0$) then the case is not suggestive of cancer ($f_2(\alpha)=0$), therefore $f_1(\alpha)=0 \Rightarrow$

$f_2(\alpha)=0$. The last two statements are equivalent to $f_2(\alpha) \geq f_1(\alpha)$ and $f_1(\alpha) \leq f_2(\alpha)$, respectively for case α. Let $E^+_{n,1}$, is a set of α sequences from E_n, such that $f_1(\alpha)=1$ (biopsy positive cases). Similarly, $E^+_{n,2}$ is a set of α sequences from E_n, such that $f_2(\alpha)=1$ (cancer positive cases). Observe, that the nested property formally means that $E^+_{n2} \subseteq E^+_{n1}$ (for all cases suggestive of cancer, biopsy should be recommended) and $f_2(\alpha) \geq f_1(\alpha)$ for all $\in E_n$.

The previous two interrelated sub-problems P1 and P2 can be formulated as a **restoration problem of two nested monotone Boolean functions** f_1 and f_2. A medical expert was presented with the ideas of monotonicity and nested functions as above and he felt comfortable with the idea of using nested monotone Boolean functions. Moreover, the dialogue, that followed, confirmed the validity of this assumption. Similarly, the function $x_2=\psi(y_1, y_2, y_3, y_4, y_5)$ for x_2 ("Shape and density of calcification") was confirmed to be a monotone Boolean function. A Boolean function is a compact presentation of the set of diagnostic rules. A Boolean discriminant function can be presented in the form of a set of logical "IF-THEN" rules, but it is not necessary that these rules stand for a single tree as in the decision tree method. A Boolean function can produce a diagnostic discriminant function, which cannot be produced by the decision tree method. For example the Biopsy Sub-Problem is stated:

$$f_1(x) = x_2x_4 \vee x_1x_2 \vee x_1x_4 \vee x_3 \vee x_5 \quad (1)$$

This formula is read as follows IF (x_2 AND x_4) OR (x_1 AND x_2) OR (x_1 AND x_4) OR (x_3) OR (x_5) THEN Biopsy is recommended. In medical terms this translates as:

IF (shape and density of calcifications suggests cancer AND comparison with previous examination suggests cancer) OR (the number and the volume occupied by calcifications suggests cancer AND shape and density of calcifications suggests cancer) OR (the number and the volume occupied by calcifications suggests cancer AND comparison with previous examination suggests cancer) OR (ductal orientation suggests cancer) OR (associated findings suggests cancer) THEN Biopsy is recommended.

The major steps in rule extraction from a medical expert are following: (1) develop a hierarchy of concepts and present them as a set of monotone Boolean functions, (2) restore each of these functions with a minimal sequence of questions to an expert, (3) combine discovered functions into a complete diagnostic function and (4) present the complete function as a traditional set of simple diagnostic rules: *If A and B and…F then Z.* Step (2 consists in the following substeps:

2.1. Expert confirms monotonicity and 'nesting" properties or redefine features to confirm these properties;

2.2. Test expert opinion about monotonicity and "nesting" properties against cases from database;2.3. Analyze database cases that violate monotonicity and "nesting", reject wrong cases;

2.4. Infer function values for derivable cases without asking an expert, but using database of cases, monotonicity and nesting properties

Interview an expert using minimal sequence of questions to completely infer a diagnostic function using monotonicity and nesting properties

The **minimal dynamic sequence of questions** is based on fundamental Hansel lemma [Hansel, 1966, Kovalerchuk et al, 1996]. We omit a detailed description of the specific mathematical steps. They can be found in [Kovalerchuk et al, 1996]. The general idea of these steps is given using an example of the interactive session in table 1. A minimal sequence of questions means that we reach the minimum of Shannon Function [Hansel, 1966, Kovalerchuk et al, 1996], i.e., a minimal number of questions is required to restore the most complex monotone Boolean function with n arguments. This sequence is not a sequence written in advance. It depends on previous answers of a medical expert, therefore each subsequent question is defined **dynamically**. Table 1 illustrates this. Columns 2, 3 and 4 present values of above defined functions f_1, f_2 and ψ (see section 3.1). We omit a restoration of function $\varphi(w_1, w_2, w_3)$ because few questions are needed to restore this function, but the general scheme is the same as for f_1, f_2 and ψ with consideration of all binary triples such as (010), (110) and so on. In table 1 the first question is: "Does the sequence (01100) represent a case requiring a biopsy?" Here, $x_1=0$ and $(01100)=(x_1, x_2, x_3, x_4, x_5)$. If the answer is "yes" (1), then the next question will be about biopsy for the case (01010). If answer is "No" (0), then the next question will be about biopsy for (11100). This sequence of questions is not accidental. As mentioned above it is inferred from Hansel lemma All 32 possible cases with five binary features (x_1, x_2, x_3, x_4, x_5) are presented in column 1 in table 1. They are grouped and the groups are called Hansel chains. The sequence of

chains begins from the shortest chain #1 --(01100) and (11100) . This chain consists of two ordered cases, (01100) < (11100) for five binary features. Then largest chain #10 consists of 6 ordered cases: (00000) < (00001) <(00011) < (00111) < (01111) < (11111). Similarly the cases are ordered as vectors in each chain.

To construct chains presented in table 1 (with five dimensions like x_1, x_2, x_3, x_4, x_5 or y_1, y_2, y_3, y_4, y_5) a sequential process is used. First, all 1-dimensional chains (in E_1) are generated and then they are used to generate chains of higher dimensions up to dimension five. Each step of chain generation consists of using current i–dimensional chains to generate (i+1) dimensional chains. The generation of chains for the next dimension (i+1) is a five–step "**clone-grow-cut-add**" process. We **clone** an i-dimensional chain, e.g., having 1-dimensional chain (0) < (1) we produce its copy: (0) < (01). Then we **grow** these chains adding the second dimension, but differently:

Chain 1: (00) < (01) Chain 2: (10) < (11) .

Here 0 is added to the left of both cases in chain 1 and 1 is added to the both cases in chain 2.
Next we **cut** the head case (11) from chain 2 and **add** it as a head to chain 1 producing two 2-dimensional Hansel chains: New chain 1--(00) < (01) < (11) and New chain 2--(10).

This process continues and stops in fifth dimension for $<x_1, x_2, x_3, x_4, x_5>$ and $<y_1, y_2, y_3, y_4, y_5>$. Table 1 presents result of this process. The chains are numbered there from 1 to 10 and each case has its number in the chain, e.g., #1.2 means the second case in the first chain. Asterisks in columns 2,3 and 4 mark answers obtained from an expert, e.g., 1* for case (01100) in column 3 means that the expert answered "yes". The remaining answers for the same chain in column 3 are automatically obtained using monotonicity. The value $f_1(01100)=1$ for case #1.1 is extended for cases #1.2, #6.3. and #7.3 in this way. Similarly values of the third monotone Boolean functions ψ are computed using the table 1. (The attributes in the sequence (10010) are interpreted as y_1, y_2, y_3, y_4, y_5 instead of x_1, x_2, x_3, x_4, x_5 used for f_1 and f_2. The Hansel chains are the same as long as the number of attributes is the same five in this case).

Column 5 and 6 list cases for extending functions' values without asking an expert. Column 5 is for extending functions' values from 1 to 1 and column 6 is for extending them from 0 to 0. If an expert were to give an answer opposite ($f_1(01100)=0$) to that presented in table 1 for function f_1 and case #1.1 (01100) then this 0 value could be extended in column 2 for cases #7.1 (00100) and #8.1 (01000). These cases are listed in column 6 for case (01100). There is no need to ask an expert about cases #7.1 (00100) and #8.1 (01000). Monotonicity provides the answer. The negative answer $f_1(01100)=0$ can not be extended for $f_1(11100)$. An expert should be queried regarding $f_1(11100)$. If his/her answer is negative $f_1(11100)=0$ then this value can be extended for cases #5.1 and # 3.1 listed in column 6 for case #1.2. Relying on monotonicity, the value of f_1 for them will also be 0.

The total number of cases with asterisk (*) in column 1 is equal to 13, for columns 3 and 4 they are respectively 13 and 12. These numbers show that 13 questions are needed to restore each of f_1 and f_2 as functions of x_1, x_2, x_3, x_4, x_5 and 12 questions are needed to restore as a function of y_1, y_2, y_3, y_4, y_5. This is only 37.5% of 32 possible questions and 60% of a possible maximum generated by Hansel lemma. Full restoration of either one of the functions f_1 and f_2 with 11 arguments (see section 3.1) without any optimization of the interview process would have required up to 2^{11}=2,048 calls (membership inquires) to the medical expert. Note that practically all studies in breast cancer computer-aided diagnostic systems derive diagnostic rules using significantly less than 1,000 cases. However, according to the Hansel lemma and under the assumption of monotony, an optimal (i.e. a minimal) dialogue for restoring a monotone Boolean would require at most: $\binom{11}{5}+\binom{11}{6} = 2 \times 462 = 924$, calls to a medical expert. This new value is 2.36 times smaller than the previous upper limit of 2,048 calls. However, even this upper limit of 924 calls can be reduced further. The hierarchy presented in fig. 5 reduces the maximum number of questions needed to restore Monotone Boolean functions of 11 binary variables to 72 questions (non-deterministic questioning) and to 46 using Hansel lemma. The actual number of questions asked was about 40, including both nested functions (cancer and biopsy) described in section 5, (i.e., about 20 questions per function).

4. DISCOVERING DIAGNOSTIC RULES FROM DATA BASE

The next task is the discovery of rules from data. This study was accomplished using an extended set of features. A set of features listed in section 3.1 was extended with two features: *Le Gal type* and *density of parenchyma* with the following diagnostic classes: "malignant", "benign", "high risk of malignancy". We extracted several dozen diagnostic rules that were statistically significant on the 0.01, 0.05 and 0.1 levels (F-criterion). The total accuracy of diagnosis is 86%. Incorrect diagnoses were obtained in 14% of diagnosed cases. The false-negative rate was 5.2% and the false-positive rate was 8.9%. Some of the rules are shown in table 2. This table presents examples of discovered rules with their statistical significance. Figure 1 presents results for another selection criterion: level of conditional probability. We studied Neural Network software had given 100% accuracy on training data, but for the Round-Robin test, the total accuracy fell to 66%. The main reason for this low accuracy is that Neural Networks (NN) do not evaluate the statistical significance of the perfect performance (100%) on training data. Poor results (76% on training data test) were also obtained with Linear Discriminant Analysis. The Decision Tree approach performed with accuracy of 76%-82% on training data. This is worse than what we obtained for the MMDR method with the much more difficult Round-Robin test (Fig. 1). The very important false-negative rate was 3-8 cases (MMDR), 8-9 cases (Decision Tree), 19 cases (Linear Discriminant Analysis) and 26 cases (NN). In these experiments, rule-based methods (MMDR and decision trees) outperformed other methods. Note also that only MMDR and decision trees produce diagnostic rules. These rules make a computer-aided diagnostic decision process visible, transparent to radiologists. With these methods radiologists can control and evaluate the decision making process. Linear discriminant analysis gives an equation, which separates benign and malignant classes. For example, 0.0670x1-0.9653x2+… represents a case. How would one interpret a weighted number of calcifications/cm2 (0.0670x1) plus a weighted volume (cm3), i.e., 0.9653x2 ? There is no direct medical sense in this arithmetic. It is hard to integrate this data mining result with knowledge management results obtained from experts.

5. RULES EXTRACTED FROM EXPERT

5.1. Examples of Diagnostic Rules exttracted using BEM method

EXPERT RULE (ER1):

IF NUMBER of calcifications per cm^2 (w_1) is large
AND TOTAL number of calcifications (w_3) is large
AND irregularity in SHAPE of individual calcifications is marked

THEN *suspicious for malignancy*

EXPERT RULE 2 (ER2):

IF NUMBER of calcifications per cm^2 (w1) large
AND TOTAL number of calcifications is large (w_3)
AND variation in SIZE of calcifications (y_3) is marked
AND VARIATION in Density of calcifications (y_4) is marked
AND DENSITY of calcification (y_5) is marked

THEN *suspicious for malignancy.*

EXPERT RULE 3 (ER3):

IF (SHAPE and density of calcifications are positive for cancer
AND Comparison with previous examination is positive for cancer)
OR (the number and the VOLUME occupied by calcifications are positive for cancer
AND SHAPE and density of calcifications are positive for cancer)
OR (the number and the VOLUME occupied by calcifications are positive for cancer AND comparison with previous examination is positive for cancer)
OR (DUCTAL orientation is positive for cancer OR associated FINDINGS are positive for cancer)

THEN *Biopsy is recommended.*

Other extracted rules in formal notation. MAL stands for suspicious for malignancy.

IF $w_2\&y_1$	THEN MAL;	IF $w_2\&y_2$	THEN MAL;
IF $w_2\&y\&_3\&y_4\&y_5$	THEN MAL;	IF $w_1\&w_3\&y_2$	THEN MAL ;
IF $w_1\&w_3\&x_5$	THEN MAL.		

Table 1. Dynamic Sequence of interview of an expert

Case	f1 biopsy	f2 Cancer	ψ shape and density of calcification	Monotone extension		Chain #	Case #
				1→ 1	0→ 0		
1	2	3	4	5	6	7	8
(01100)	1*	1*	1*	1.2;6.3;7.3	7.1;8.1	Chain 1	1.1
(11100)	1	1	1	6.4;7.4	5.1;3.1		1.2
(01010)	1*	0*	1*	2.2;6.3;8.3	6.1;8.1	Chain 2	2.1
(11010)	1	1*	1	6.4;8.4	3.1;6.1		2.2
(11000)	1*	1*	1*	3.2	8.1;9.1	Chain 3	3.1
(11001)	1	1	1	7.4;8.4	8.2;9.2		3.2
(10010)	1*	0*	1*	4.2;9.3	6.1;9.1	Chain 4	4.1
(10110)	1	1*	1	6.4;9.4	6.2;5.1		4.2
(10100)	1*	1*	1*	5.2	7.1;9.1	Chain 5	5.1
(10101)	1	1	1	7.4;9.4	7.2;9.2		5.2
(00010)	0*	0	0*	6.2;10.3	10.1	Chain 6	6.1
(00110)	1*	1*	0*	6.3;10.4	7.1		6.2
(01110)	1	1	1	6.4;10.5			6.3
(11110)	1	1	1	10.6			6.4
(00100)	1*	1*	0*	7.2;10.4	10.1	Chain 7	7.1
(00101)	1	1	0*	7.3;10.4	10.2		7.2
(01101)	1	1	1*	7.4;10.5	8.2;10.2		7.3
(11101)	1	1	1	5.6			7.4
(01000)	0*	0	1*	8.2	10.1	Chain 8	8.1
(01001)	1*	1*	1	8.3	10.2		8.2
(01011)	1	1	1	8.4	10.3		8.3
(11011)	1	1	1	10.6	9.3		8.4
(10000)	0*	0	1*	9.2	10.1	Chain 9	9.1
(10001)	1*	1*	1	9.3	10.2		9.2
(10011)	1	1	1	9.4	10.3		9.3
(10111)	1	1	1	10.6	10.4		9.4
(00000)	0	0	0	10.2		Chain 10	10.1
(00001)	1*	0*	0	10.3			10.2
(00011)	1	1*	0	10.4			10.3
(00111)	1	1	1	10.5			10.4
(01111)	1	1	1	10.6			10.5
(11111)	1	1	1				10.6
Total Calls	13	13	12				

Table 2. Examples of extracted diagnostic rules

Diagnostic rule	F-criterion for features		total significance of F-criterion			Accuracy for test cases (%)
			0.01	0.05	0.1	
IF NUMber of calcifications per cm2 is between 10 and 20 AND VOLume > 5 cm3 THEN Malignant	NUM VOL	0.0029 0.0040	+ +	+ +	+ +	93.3
IF TOTal number of calcifications >30 AND VOLume > 5 cm3 AND DENSITY of calcifications is moderate THEN Malignant	TOT VOL DEN	0.0229 0.0124 0.0325	- - -	+ + +	+ + +	100.0
IF VARiation in shape of calcifications is marked AND NUMber of calcifications is between 10 and 20 AND IRRegularity in shape of calcifications is moderate THEN Malignant	VAR NUM IRR	0.0044 0.0039 0.0254	+ + -	+ + +	+ + +	100.0
IF variation in SIZE of calcifications is moderate AND Variation in SHAPE of calcifications is mild AND IRRegularity in shape of calcifications is mild THEN Benign	SIZE SHAPE IRR	0.0150 0.0114 0.0878	- - -	+ + -	+ + +	92.86

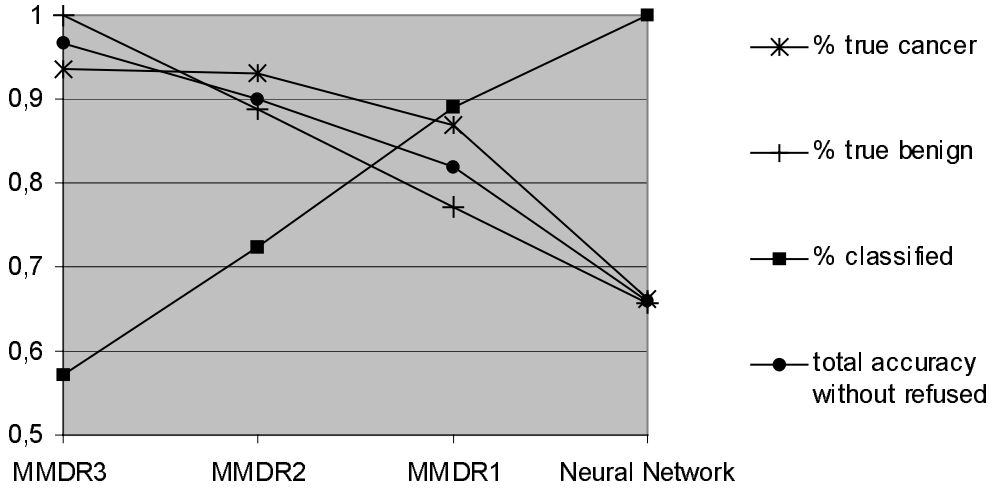

Figure 1. Performance of methods (Round-Robin test)

5.2. Rule extraction through monotone Boolean functions

Boolean expressions for shape and density of calcification $x_2 = \psi(y_1, y_2, y_3, y_4, y_5)$ were obtained from the information depicted in table 1 (columns 1 and 4) with the following steps:

(i) Find all the maximal lower units for all chains as elementary conjunctions;

(ii) Exclude the redundant terms (conjunctions) from the end formula. See expression (2) below.

Thus, from table 1 (columns 2, 4) we obtained

$$x_2=\psi(y_1, y_2, y_3, y_4, y_5) =y_1y_2y_2y_3\vee y_2y_4\vee y_1y_3\vee y_1y_4\vee y_2y_3y_4\vee y_2y_3y_5\vee y_2\vee y_1\vee y_3y_4y_5$$

and then we simplified it to $y_2\vee y_1\vee y_3y_4y_5$. As above, from columns 2 and 3 in table 1 we obtained the initial components of the target functions of x_1, x_2, x_3, x_4, x_5 for the biopsy sub-problem as follows:

$$f_1(x) = x_2x_3\vee x_2x_4\vee x_1x_2\vee x_1x_4\vee x_1x_3\vee x_3x_4\vee x_3\vee x_2x_5\vee x_1x_5\vee x_5,$$

and for the cancer sub-problem to be defined as:

$$f_2(x) = x_2x_3\vee x_1x_2x_4\vee x_1x_2\vee x_1x_3x_4\vee x_1x_3\vee x_3x_4\vee x_3\vee x_2x_5\vee x_1x_5\vee x_4x_5.$$

The simplification of these disjunctive normal form (DNF) expressions allowed us to exclude some redundant conjunctions. For instance, in x_2 the term y_1y_4 is not necessary, because y_1 covers it. Thus, the right hand side of expressions (1) to (4) forms the minimal disjunctive normal form (DNF).

Using this technique 16 rules were extracted for the diagnostic class "suspicious for malignancy" and 13 rules for the class "biopsy" (see formulas (5) and (6) for mathematical presentation).

All these rules are obtained from formula (6) presented below.

Similarly, for the second sub-problem (highly suspicious for cancer) the function that we found was:

$$f_2(x) = x_1x_2\vee x_3\vee (x_2\vee x_1\vee x_4)x_5 \qquad (2)$$

Regarding the second level of the hierarchy (which recall has 11 binary features) we interactively constructed the following functions (interpretation of the features is presented below):

$$x_1 = \nu(w_1, w_2, w_3) = w_2 \vee w_1 w_3 \quad (3)$$

and $$x_2 = \psi(y_1, y_2, y_3, y_4, y_5) = y_1 \vee y_2 \vee y_3 y_4 y_5 \quad (4)$$

By combining the functions in (1)-(4) we obtained the formulas of all 11 features for biopsy:

$$f_1(x) = (y_2 \vee y_1 \vee y_3 y_4 y_5) x_4 \vee (w_2 \vee w_1 w_3)(y_2 \vee y_1 \vee y_3 y_4 y_5) \vee (w_2 \vee w_1 w_3) x_4 \vee x_3 \vee x_5 \quad (5)$$

and for suspicious for cancer:

$$f_2(x) = x_1 x_2 \vee x_3 \vee (x_2 \vee x_1 \vee x_4) x_5 = (w_2 \vee w_1 w_3)(y_1 \vee y_2 \vee y_3 y_4 y_5) \vee x_3 \vee (y_1 \vee y_2 \vee y_3 y_4 y_5) \vee (w_2 \vee w_1 w_3 \vee x_4) x_5 \quad (6)$$

6. COMPARISON OF DATA-BASED AND EXPERT DIAGNOSTIC RULES

Below we compare some rules extracted using data mining algorithms and by interviewing the radiologist. The rule DBR1 was extracted from the data: *IF NUMber of calcifications per cm^2 (w_1) is between 10 and 20 AND VOLume (w_2)>5 cm^3*
THEN Malignant

The closest expert rule is ER1: *IF NUMber of calcifications per cm^2 (w_1) large AND TOTal number of calcifications (w_3) is large AND irregularity in SHAPE of individual calcifications (y_1) is marked THEN Malignant*

There is no rule DBR1 among the expert rules, but this rule is statistically significant (0.01, F-criterion). Rule DBR1 should be tested by the radiologist and included in diagnostic knowledge base after expert's verification. The same verification procedure should be done for ER1. This rule should be analyzed against database of real cases. This analysis may lead to conclusion that the database is not sufficient and rule DB1 should be extracted from the extended database. Also, the radiologist can conclude that the feature set is not sufficient to incorporate rule DBR1 in to his knowledge base. This kind of analysis is not possible for Linear Discriminant analysis or Neural Networks. We also use fuzzy logic to clarify the meaning of such concepts as "Total number of calcifications (w_3) is *large*" during the comaparing rules. The reliability of the expert radiologist was tested against actual cases, classified into three categories: (1)"High probability of Cancer, Biopsy is necessary" (or CB), (2)"Low probability of cancer, probably Benign but Biopsy/short term follow-up is necessary" (or BB) and (3)"Benign, biopsy is not necessary" (or BO). These cases were selected from screening cases recalled for magnification views of calcifications. For the CB and BB cases, pathology reports of biopsies confirmed the diagnosis while a two-year follow-up had been used to confirm the benign status of BO. The expert's diagnosis was in full agreement with his extracted diagnostic rules for 60% of the cases and for 40% of the cases the expert asked for more information than that given in the extracted rule. Diring the inteview the expert stated that that he had cases with the same combination of 11 features but with different diagnosis. This suggests that the feature set should be extended to adequately cover complicated cases. Restoration of Monotone Boolean functions allowed us to identify this need. This is one of the useful outputs from these functions.

The following rule DBR2 was extracted from the database:
IF variation in SIZE of calcifications is moderate AND variation in SHAPE of calcifications is mild AND IRRegularity in shape of calcifications is mild THEN Benign.

This rule is confirmed by the database of actual cases using the Round-Robin test. We extracted from this database all cases for which this rule is applicable, i.e., cases where the variation in SIZE of calcifications is moderate; variation in SHAPE of calcifications is mild and IRRegularity in shape of calcifications is mild. For 92.86% of these cases the rule is accurate. The expert also has a rule with these premises, but the expert rule includes two extra premises: ductal orientation is not present and there are no associated findings (see formula (6)). This suggests that the database should be extended to determine which rule is correct.

Radiologists Comments regarding Rules extracted from Database

DB RULE 1:

IF TOTAL number of calcifications >30
AND VOLUME >5 cm^3
AND ***DENSITY of calcifications is moderate*** *THEN* ***Malignant.***

F-criterion—significant for 0.05. Accuracy of diagnosis for test cases --100%.

Radiologist's comment—This rule might have promise, but I would consider it risky.

DB RULE 2:

IF VARIATION in shape of calcifications is marked
AND NUMBER of calcifications is between 10 and 20
AND IRREGULARITY in shape of calcifications is moderate

THEN ***Malignant.***

F-criterion—significant for 0.05. Accuracy of diagnosis for test cases -- 100%.

Radiologist's comment—I would trust this rule.

DB RULE 3:

IF variation in SIZE of calcifications is moderate
AND variation in SHAPE of calcifications is mild
AND IRREGULARITY in shape of calcifications is mild

THEN Benign.

F-criterion—significant for 0.05. Accuracy of diagnosis for test cases -- 92.86%.

Radiologist's comment—I would trust this rule.

.

7. DISCUSSION AND CONCLUDING REMARKS

The study has demonstrated how integrated consistent data mining can acquire a set of logical diagnostic rules for intergrated DM/KM computer-aided diagnostic systems. Consistency avoids contradiction between rules generated using data mining software, rules used by an experienced radiologist, and a database of pathologically confirmed cases. Two major problems: (P1) to find contradiction between diagnostic rules and (P2) to eliminate contradiction are identified. Two complimentary intelligent technologies for extracting rules and discoivering their contradiction have been applied. The first technique is based on discovering statistically significant logical diagnostic rules. The second technique is based on the restoration of a monotone Boolean function to generate a minimal dynamic sequence of questions to an expert. The results of this mutual verification of expert and data-driven rules demonstrate feasibility of the approach for designing integrated consistent DM and KM systems.

8. REFERENCES

1. Dhar V, Stein R: *Intelligent Decision Support Methods*. Prentice Hall, NJ, 1997.
1. Hansel G: Sur le nombre des fonctions Boolenes monotones den variables. *C.R. Acad. Sci*. Paris, 262(20):1088-1090, 1966.
2. Kovalerchuk B., Vityaev E: *Data Mining in Finance: Advances in Relational and Hybrid methods*, Kluwer Academic Publishers, 2000, p.308.
3. Kovalerchuk, B., E. Triantaphyllou, A. Despande, and E. Vityaev, Interactive Learning of Monotone Boolean Function. *Information Sciences*, Vol. 94, issue 1-4, 1996, pp. 87-118.
4. Kovalerchuk, B., Vityaev, E., Ruiz, J. Consistent knowledge discovery in medical diagnosis, *IEEE Engineering in Medicine and Biology*, vol. 19, n. 4, July/August 2000, pp. 26-37.
5. Krantz DH, Luce RD, Suppes P, Tversky A: *Foundations of Measurement*, v.1-3, Acad. Press, NY, London. 1971, 1989, 1990.
6. Mitchell T: *Machine Learning*. NY, McCraw Hill, 1997
7. Russel S, Norvig P: *Artificial Intelligence. A Modern Approach*, Prentice Hall, 1995
8. Vityaev EE: Semantic approach to knowledge base development: Semantic probabilistic inference. *Computational Systems* 146: 19-49, Novosibirsk, 1992 (in Russian).
9. Vityaev EE, Moskvitin AA: Introduction to discovery theory: Discovery software system. *Computational Systems*, 148: 117-163, Novosibirsk, 1993 (in Russian).