Empirical Theories Discovery based on the Measurement Theory

Vityaev E.E.[1]*, Kovalerchuk B.Y.[2]

[1] Sobolev Institute of Mathematics SB RAS, Acad. Koptyug prospect 4, Novosibirsk, 630090, Russia.

[2] Boris Kovalerchuk, Computer Science Department, Central Washington University, Ellensburg, WA, 98926-7520, USA.

* Corresponding author. E-mail: vityaev@math.nsc.ru

Summary

The purpose of this work is to analise the cognitive process of the domain theories as it described in the Measurement Theorry and to develop a computational Machine Learning approach, which realise it. As a result the Relational Data Mining approach, proposed by the authors in the preceding books, is developed. We present this approach, as implementation of the cognitive process as it perceived in the Measurement Thoery. In the first part of the paper we analise the cognitive process. In the second we persent the theory and method of the logically most powerful empirical theory discovery. It is based on the notion of "law-like" rules, which satisfy all the properties of laws of nature: generality, simplicity, maximum refutability and minimum number of parameters. This notion is defined for deterministic and probabilistic cases. Based on this method the system "Discovery" is developed. This system was succesfully applied for solving of many practical tasks.

Keywords: *Machine Learning, Knowledge Discovery, Data Mining, Scientific Discovery, Discovery Science, Domain Theory, Theory Formation*

1. Introduction.

A variety of Machine Learning methods have been developed, but their relation to cognitive process remains unclear. What is the cognitive process? Let us consider the cognitive process as it conceived in the Measurement Theory.

The Representative Measurement Theory (RMT) was originated by P. Suppes and other scientists at Stanford University [Scott, Suppes, 1958; Suppes, Zines, 1963, Krantz at al., 1971, 1981, 1990, Narens, 1985]. The numerical representations of the values and laws of nature are investigated in this theory. It was shown that numerical representations of the values and laws are only *numerical codes* of the *algebraic structures* representing the operational properties of these values and laws. Thus, the algebraic structures are *primary representations* of values and laws. The main postulates and results of the Measurement Theory are the following [Krantz at el, 1971, 1981, 1990]:

- numerical representations of quantities and laws of nature are determined by the set of axioms for corresponding empirical systems – algebraic systems with some sets of relations and operations;
- this numerical representations are unique up to some sets of allowable transformations (such as a change of measurements units);
- all physical attributes may be embedded into the structure of physical quantities;
- physical laws are simple because of procedure of simultaneous scaling of all attributes involved in the law (there is no machine learning method, which may perform such discovering of laws);
- the same axiomatic approach is also applicable not only for physical attributes and laws, but also for many other attributes from other domains (such as psychology) using polynomial and other representations;

It follows from the Representative Measurement Theory that a cognition process for some domain may be performed in the following steps:

I. identify the domain theory as a set of its attributes and laws;

II. determine a set of relations and operations which constitute their operational sense of the domain attributes and laws;

III. discover all sets of axioms which are fulfilled for these relations and operations;

IV. determine numerical representations of all attributes and laws using theorems of the Representative Measurement Theory. These theorems determine the numerical representations of attributes and laws for corresponding sets of axioms;

V. perform a simultaneous scaling procedure for all attributes involved in the discovered laws;

VI. determine the structure of all (rescaled) quantities interconnected by the laws for the explored domain theory.

This description of the cognition process is one of the most elaborated and formalized in the contemporary science. However this cognition process has the following limitations:

(a) there is no a general method for determining the sets of relations and operations and describing the operational sense of attributes and laws;

(b) there is no general method for axiom systems of empirical systems discovery in the noise conditions;

(c) the Measurement Theory theorems on existence of numerical representations are known only for some axiom systems. For the most axiom systems there is no such theorems;

(d) there is no general method of simultaneous scaling of all attributes involved in the law.

These limitations severely restrict possibilities to implement this cognition process computationally.

Let us analyze this cognition process from machine learning viewpoint. Let us divide this cognitive process in two steps:

Step 1: determine the set of relations and operations and discover all axiom systems for every attribute and law;

Step 2: determine numerical representations of all attributes and laws using the theorems from Measurement Theory. For any discovered law perform a simultaneous scaling procedure for all attributes involved in the law. Determine a structure of all (rescaled) attributes interconnected by laws for the explored domain theory.

Step 1 implements a Logical Empirical Theory discovery (LET) of the explored domain. The step 2 implements the Quantitative Empirical Theory (QET) discovery based on the results of Measurement Theory. According to the methodology of the Measurement Theory the cognition processes need to begin from discovering Logical Empirical Theory. The LET represents the qualitative description of the explored domain. A *transformation* of the domain theory from the qualitative state to quantitative must be performed using Measurement Theory. The process of a domain theory transformation reflects the history of the domain theory growth. This important aspect is missed in the standard Machine Learning approach, which does not use deep results from the Measurement Theory.

Let us abandon ourselves in the domain theory exploration by the step 1, that is by the Logical Empirical Theories. There are some reasons for that:

- Step 2 need to be performed using Measurement Theory, not by the Machine Learning methods;
- Step 2 follows the historical traditions – it is very convenient for human to use numbers for attributes and laws. The pure operational and algebraic representation is unacceptable for humans. Nevertheless, nowadays we have instruments to operate with the Logical Empirical Theories using, for example, logical programming methods;

- Step 2 creates some limitations for to the cognitive process: there are attributes that have no natural numerical representations. Such attributes as structural attributes, partial orders, lattices, graphs, some tests results, preference relation and so on. The same limitation takes place for such laws as diagnosis, utility functions, trading figures, psychological tests that are not always have an appropriate numerical representation;
- Step 2 produces the Domain Theory as a Quantitative Empirical Theory represented by the structure of attributes interconnected by the laws. This representation is well elaborated. The Domain Theory as Logical Empirical Theory has no yet a convenient representation. But some representations are developed nowadays, for example a Logical Empirical Theory may be represented as an expert system or as a logic program.

We consider the Logical Empirical Theories as more adequate and modern way for a domain theory representation. In accordance with this idea we developed a Relational Data Mining (RDM) approach [Kovalerchuk, B., Vityaev, E., 2000] to the Logical Empirical Theories discovery. It integrates the achievements of the Measurement Theory and objectives of the Machine Learning and KDD&DM. RDM approach intends to overcome limitations of the Measurement Theory and Machine Learning identified above. Specifically according the Measurement Theory any numerical type of data can be transformed into the relational form with complete preservation of the relevant properties of numeric data type but it is not a systematic part of modern Machine Leaning approach. The cognitive process in Relational Data Mining consists in the following actions:

(a) transform all empirically interpretable information from data into a many-sorted empirical system – an algebraic systems with empirically interpretable sets of relations and operations;

(b) discover a Logical Empirical Theory for the many-sorted empirical system with noise conditions using a specially developed RDM method "Discovery", which discovers regularities as logical expressions in the first-order logic with probabilistic estimates;

(c) use a hierarchy of data types to speed up the search for regularities. The search begins with testing rules based on the properties from the weaker scales and finishes with properties of stronger scales. The number of search computations for weaker scales is smaller than for stronger scales;

The possibility to implement actions a) and b) was demonstrated in [Kovalerchuk, B., Vityaev, E., 2000] for a variety of data types as matrices of comparisons of pairs, and multiple comparisons, attribute-based matrices, matrices of ordered data, and matrices of closeness. These structures have been converted to many-sorted empirical systems. We argue that current Machine Learning methods utilize only a part of data type information actually presented in data. Incorporation of such information opens an enormous opportunity to enhance performance of Machine Learning methods.

For implementing action c) we developed a rather general RDM method "Discovery", which is presented in this paper. The generality of the method is partially justified by its product that is a "strongest Logical Empirical Theory" defined formally below. We prove (see paragraph 2) such strongest Logical Empirical Theory is a product of discovery by using this method. Thus it implements the cognitive process of LET discovery.

2. Representative Measurement Theory

A relational structure consists of a set of objects A and k(i)–ary relations $P_1,\ldots,P_n$ and k(j)-ary operations $\rho_1,\ldots, \rho_m$ defined on A

$$\mathbf{A} = \langle A, P_1,\ldots,P_n, \rho_1,\ldots, \rho_m \rangle$$

Each relation P_i is a Boolean function (predicate) with k(i) arguments from A and ρ_j is k(j) arguments operation on A. The relational structure $A = \langle A,S_1,\ldots,S_n \rangle$ is considered along with a numerical structure of the same type

$$\mathbf{R} = \langle R,T_1,\ldots,T_n, \sigma_1,\ldots, \sigma_m \rangle$$

Where set R is a subset of $\mathrm{Re}^{\boldsymbol{m}}$, $m \geq 1$, where $\mathrm{Re}^{\boldsymbol{m}}$ is a set of m-tuples of real numbers and each relation T_i has the same arity n_i as the corresponding relation P_i and each real-valued function σ_j has the same arity k(j) as the corresponding operation ρ_j. The relational structure **A** is interpreted as an empirical real-world system and **R** is interpreted as a numerical system designed as a numerical representation of **A**. The idea of numeric representation formalized by the notion of a homomorphism φ: $\mathbf{A} \rightarrow \mathbf{R}$.

A mapping φ: $A \rightarrow \mathrm{Re}^{\boldsymbol{m}}$ is called a (srtong) *homomorphism* iff:

$$P_i(a_1,\ldots,a_{k(i)}) \Leftrightarrow T_i(\varphi(a_1),\ldots, \varphi(a_{k(i)})),\ i = 1,\ldots,n;$$

$$\varphi(\rho_j(a_1,\ldots,a_{k(j)})) = \sigma_j(\varphi(a_1),\ldots, \varphi(a_{k(j)})),\ j = 1,\ldots,m.$$

Numerical system **R** is called a *numerical representation* of the relational structure **A** if a homomorphism φ: $\mathbf{A} \rightarrow \mathbf{R}$ exists. In the Measurement Theory the following theorems are proved: given relational structure **A** find a numerical representation **R** for **A** and prove the (existence) theorem that a homomorphism φ: $\mathbf{A} \rightarrow \mathbf{R}$ exists; given relational structure **A** and

numerical representation **R** define the set of all homomorphisms φ: **A** → **R** (uniqueness theorems).

Consequently, relational structure is represented in numerical and hence computationally tractable form with complete reservation of all properties of relational structure.

Example: A relational structure **A** = < A,P> called a semi-ordering if it satisfies the axiom:

a,b,c ∈ A (P(a,b)&P(b,c) ⇒ ∀d∈ A(P(a,d)∨P(d,c))).

Theorem: If **A** = < A,P> is a semi-ordering then there exists a function U: A → Re such that:

P(a,b) ⇔ U(a) + 1 < U(b).

3. Data types problems

Let us consider problems of the action a) of RDM cognitive process.

A data type in object-oriented programming languages is a relational structure **A** with the sets of relations and operations, which are interpretable in the domain theory. For instance, a "stock price" data type may be presented as relational structure **A** = < A; {≤, =, ≥}> with nodes A reflecting individual stock prices and edges reflecting relations between stock prices {≤, =, ≥}. So data type is a relational structure.

Implicitly, each attribute in data reflects a data type, which can take a number of possible values. These values are elements of A. For instance, attribute "date" has 365 (366) elements from 01.01.2000 to 12.31.2001. There are several meaningful relations and operations with dates: <, =, >, and middle(a,b). For instance, the operation middle(a,b) produces the middle date c=01.05.99 for inputs a=01.03.99 and b=01.07.99. It is common in attribute-value languages (AVL) that such data type as a date given as an implicit data type. Usually relations P and operations F are not expressed explicitly.

Let us examine the empirical status of data types in Machine Learning and Data Mining. To extract all empirically interpretable information from data a relational structure **A** representing a data type should be interpreted as an empirical real-world structure, i.e., A and the sets of operations and functions should be included in the domain background knowledge of the learning task. Numerical Machine Learning methods assume that any numerical standard mathematical operations (+,-,*, / and so on) can be used in algorithm despite their possible non-interpretability. In this way, a non-interpretable results may be obtained as well. Let us consider this situation in more detail for six different cases:

1. *Physical data types in physical domains*. Data contain only physical quantities and domain background knowledge of the learning task belongs to physics, where data types and measurement procedures are well developed. In this case, the measurement theory [Krantz et al, 1970, 1981, 1990] provides formalized relational structures for all quantities. The use of the Machine Learning methods is most appropriate in this case.
2. *Physical data types for non-physical problems*. Data contain physical quantities, but the domain background knowledge of the learning task is not belongs to physics. The background knowledge may belong to finance, geology, medicine, and other areas. In this case, actual data types are not known even when they represent physical quantities. If the quantity is physical, then we know the relational structure from the measurement theory. But relations of that relational structure are physically interpretable and not necessarily interpretable in domain background knowledge. We need to establish interpretation of relations and operations for the new domain. If relation isn't interpretable we need to remove it from the relational structure. For example, for many physical quantities there is interpretable operation $\circ$, matched to the formal numeric additive operation + and its properties. However, the use of the same attribute in finance, medicine, and other fields can change its data type. For instance, there is no known medical procedure on a patient

that gives us temperature t3 from two patient's temperatures t1 and t2, $t3=t1 \circ t2$. Therefore, the operation $\circ$ doesn't have interpretation in medicine. On the other hand, the relation "<" makes sense in both areas. Everyone knows the meaning of increasing temperature. It means that temperature data type in physics differs from temperature data type in medicine and we do not know exactly what the medical temperature data type is. The interpretation of the temperature in background knowledge of medicine needs to interpret the temperature in terms of metabolism. The physical temperature measured by thermometer in this case is indirect measure of the some medicine quantity expressing the speed of metabolic processes.

3. *Non-physical data types for non-physical tasks*. For non-physical quantities, data types are practically unknown. There are two sub cases:
    a. Non-numerical data types. It was demonstrated [Kovalerchuk, B., Vityaev, E., 2000] how such non-numerical data types as matrices of pairs-comparisons, and multiple-comparisons, attribute-based matrices, matrices of ordered data, and matrices of closeness may be represented in many sorted empirical systems in rather natural form by representing them in terms of the first-order logic;
    b. Numerical data types. In this case we have a measurer x(a), which produce a numeric number as a result of the measurement procedure applied to object a. The examples of measurers are psychological tests, stock market indicators, questionnaires, and physical measurers used in non-physical areas.

       Let us define a set of empirically interpretable relations and operations for the measurer x(a). For any numerical relation $T(y_1,\ldots,y_k)$ and operation σ in Re^m (Re - the set of real numbers), we may define the following empirical relation P^T on A^k and operation ρ^σ on A^m

$$P^T(a_1,\ldots,a_k) \Leftrightarrow R(x(a_1),\ldots,x(a_k));$$

$$(b = \rho^{\sigma}(a_1,\ldots,a_m)) = (x(b) = \sigma(x(a_1),\ldots,x(a_m)))$$

The measurer x obviously has an empirical interpretation, but relation P^T and operation ρ^{σ} may not. We need to find such relations T and operations σ that have empirical interpretation in domain background knowledge. The set of obtained interpretable relations will not be empty, because at least the relation $P^{=}$ has an empirical interpretation: $P^{=}(a_1,a_2) \Leftrightarrow x(a_1) = x(a_2)$.

In measurement theory, there are many sets of axioms based on just ordering and equivalence relations. Nevertheless, these sets of axioms establish strong data types. Strong data types are a result of interactions of the quantities with weak data types such as ordering and equivalence. For instance, having one weak order relation $<_y$ (for attribute y) and n equivalence relations $\approx_{x1},\ldots, \approx_{xn}$ for attributes $x_1,\ldots,x_n$, we can construct a complex relation (defined by axiom system) between y and $x_1,\ldots,x_n$ given by $G(y,x_1,\ldots,x_n) \Leftrightarrow y = f(x_1,\ldots,x_n)$, where $f(x_1,\ldots,x_n)$ is a polynomial [Krantz et al, 1971]. This is a very strong result. To construct a polynomial we need the multiplication, power and sum operations, but this operations is not defined for $x_1,\ldots,x_n$. However, relation G is equivalent to polynomial f if a certain set of axioms expressed in terms of order relation $<_y$ and equivalence relations $\approx_{x1},\ldots, \approx_{xn}$ are true for x. Ordering and equivalence relations are usually empirically interpretable in background knowledge for different domains.

4. *Nominal discrete data types*. In this case, all data are interpretable in corresponding relational structures because there is no difference between the numerical and empirical sys-

tems. All numbers are only names, and names can be easily represented as predicates with one variable.

5. *Non-quantitative and non-discrete data types*. Data contain no quantities and discrete variables, but do contain ranks, orders and other nonstandard data types. This case is similar to the case 3a. The only difference is that such data usually are made discrete using various calibrations with losing useful information.
6. *Mix of data types*. All mentioned difficulties arise in this case. To be able to work with all sorts of data type mixes, a new approach is needed. Relational Data Mining implements this approach using relational representation of data types.

4. First-order logic approaches

It is clear from previous definitions that only Machine Learning methods based on the first-order logic (FOL) may discover Logical Empirical Theory. Let us consider the existent FOL methods.

A variety of relational machine learning systems have been developed in recent years [Mitchell, 1997]. Theoretically, these systems have many advantages. In practice though, the complexity of the language must be severely restricted, reducing their applicability. For example, some systems require that the concept definition be expressed in terms of attribute-value pairs [Lebowitz, 1986; Danyluk, 1989] or only in terms of unary predicates [Hirsh, 1989; Mooney, Ourston, 1989; Katz, 1989; Shavlik, Towell, 1989; Pazzani, 1989; Sarrett, Pazzani, 1989]. The systems that allow actual relational concept definitions (e.g., OCCAM [Pazzani, 1990], IOE [Flann & Dietterich, 1989], ML-SMART [Bergadano et al., 1989]) place strong restrictions on the form of induction and the initial knowledge that is provided to the system [Pazzani, Kibler, 1992].

The major successful applications of FOL are presented in [Bratko et al., 1992; Muggleton et al., 1992; Muggleton, 1999; Bratko, 1993; Dzeroski et al., 1994; Kovalerchuk et al., 1997; Pazzani, 1997]; Dzeroski [1996], Bratko, Muggleton [1995], Muggleton [1999] and Pazzani [1997]. These methods have been successfully applied to many problems in chemistry, physics, medicine and other fields. Such tasks as mesh design, mutagenicity, and river water quality exemplifies successful applications. Domain specialists appreciate that the learned regularities are understandable directly in domain terms. Financial applications can specifically benefit from these methods. Fu [1999] noted "Lack of comprehension causes concern about the credibility of the result when neural networks are applied to risky domains, such as patient care and financial investment".

Advantages of the first-order logic (FOL) methods.

Predicate invention. Human-readable and understandable form of rules. To utilize advantages of human-readable forecasting rules logical relations (predicates) should be developed. In this way, FOL methods can produce valuable understandable rules in addition to the forecast. Using this technique a financial specialist can evaluate the performance of the forecast as well as a forecasting rule. Obviously, understandable rules have advantages over a stock market forecast without explanations. The problems of inventing predicates were considered in the previous section and in [Kovalerchuk, Vityaev, 2000].

Advantages and disadvantages of attribute-value languages (AVLs) methods and first order logic (FOL) methods (table 1). Bratko and Muggleton [1995] pointed out that existing FOL systems are relatively inefficient and have rather limited facilities for handling numerical data. The purpose of Relational Data Mining (RDM) is to overcome these limitations of current FOL methods. There are two types of numerical data in data mining: (i) the numerical target variable and (ii) numerical attributes used to describe objects and discover patterns. Traditionally FOL methods solve only classification tasks without direct operations on nu-

merical data. The "Discovery" system handles an interval forecast of numeric variables with continuous values like prices along with solving classification tasks. In addition, "Discovery" system handles numerical time series using the first-order logic technique, which is not typical for ILP and FOL applications.

Table.1. Comparison of AVL-based methods and first-order logic methods

Method	Advantages for the learning process	Disadvantages for the learning process
Methods based on attribute-value languages	**Simple**, **efficient**, and handle **noisy data**.	**Limited form** of background **knowledge**. **Lack of relations** in the concept description language.
Methods based on First Order Logic	Appropriate learning time with a large **number of training examples**. Solid **theoretical basis** (first-order logic, logic programming). **Flexible** form of background knowledge, problem representation, and problem-specific constraints. **Understandable representation** of background knowledge, and relations between examples.	Inappropriate learning time with a large **number of arguments** in the relations. **Weak facilities** for processing **numerical data.**

Background knowledge. Knowledge Representation is an important and informal initial step in Relational Data Mining. In attribute-based methods, the attribute form of data actually dictates the form of knowledge representation. Relational Data Mining has more options for knowledge representation. For example, attribute-based stock market information such as stock prices, indexes, and volume of trading should be transformed into the first-order logic form. This knowledge includes much more than only values of attributes. There are many ways to represent knowledge in the first-order logic language. Data Mining algorithms may work too long to "dig" relevant information or even may produce inappropriate rules. Introducing data types [Flash et al., 1998] and concepts of representative measurement theory [Krantz et all, 1971, 1989, 1990, Narens, 1985; Pfanzagl, 1968] into the knowledge represen-

tation process helps to address this representation problem. In fact the measurement theory developed a wide set of data types, which cover data types used in [Flash et al., 1998]. FOL systems have a mechanism to represent background knowledge in human-readable and understandable form.

Hybridizing the logical data mining methods with a probabilistic approach (see definition of "probabilistic laws" section 8). This is done by introducing probabilities over logical formulas [Carnap, 1962; Fenstad, 1967; Vityaev E. 1983; Halpern, 1990, Vityaev, Moskvitin, 1993; Muggleton, 1994, Vityaev et al, 1995; Kovalerchuk, Vityaev, 1998, 2000, Koller, Pfeffer, 1997]. The is one of the few Hybrid Probabilistic Relational Data Mining methods developed and applied to financial data [Kovalerchuk, Vityaev, 1998, 2000; Vityaev et al, 1995; Vityaev, Moskvitin, 1993; Vityaev E., 1983]. The "Discovery" system has been applied to predict SP500C time series and to develop a trading strategy. This method outperformed several other strategies in simulated trading [Kovalerchuk, Vityaev, 1998, 2000].

Statistical significance. Traditionally, FOL methods were pure deterministic techniques, which originated in logic programming. There are well-known problems with deterministic methods - handling data with a significant level of noise. This is especially important for financial data, which typically have a very high level of noise. On the other hand, RDM should handle imperfect (noisy) data and in particular imperfect numerical data. Statistical significance is another challenge for deterministic methods. Statistically significant rules have an advantage in comparison with rules tested only for their performance on training and test data [Mitchell, 1997]. Training and testing data can be too limited and/or not representative. If rules rely only on them then there are more chances that these rules will not deliver a correct forecast on other data. This is a hard problem for any data mining method and especially for deterministic methods including deterministic ILP. Intensive studies are being conducted for incorporating a probabilistic mechanism into ILP [Muggleton, 1994].

Search space for hypotheses. It is well known that the general problem of rule generating and testing is NP-complete [Hyafil, Rivest, 1976]. Therefore, the discussion above is closely related to the following questions. What determines the number of rules? When do we stop generating rules? The number of hypotheses is another important parameter. It has already been mentioned that RDM with first-order rules allows one to express naturally a large variety of general hypotheses. These more general rules can be used for classification problems as well as for an interval forecast of a continuous variable. FOL algorithms face exponential growth in the number of combinations of predicates to be tested. A mechanism to decrease this set of combinations is needed. To addresses these issues we propose a data type system and the representative measurement theory approach. Type systems and measurement theory approaches provide better ways to generate only meaningful hypotheses using syntactic information. A probabilistic approach also naturally addresses knowledge discovery in situations with incomplete or incorrect domain knowledge. In this way the properties of individual examples are not generalized beyond the limits of statistically significant rules.

Algorithms FOIL, FOCL and "Discovery". Algorithm FOIL [Quinlan, 1989; Quinlan, 1990] learns constant-free Horn clauses, a useful subset of first-order predicate calculus. Later FOIL was extended to use a variety of types of background knowledge to increase the class of problems that can be solved, to decrease the hypothesis space explored, and to increase the accuracy of learned rules.

Algorithm FOCL [Pazzani, Kibler, 1992] extends FOIL. FOCL uses first-order logic and FOIL's information-based optimality metric in combination with background knowledge. This is reflected in its full name -- First Order Combined Learner. FOCL has been tested on a variety of problems [Pazzani, 1997] that includes a domain theory describing when a student loan is required to be repaid [Pazzani & Brunk, 1990].

As we mentioned the general problem of rule generating and testing is NP-complete [Hyafil, Rivest, 1976]. Therefore, we face the problem of designing NP-complete algorithms. There are several related questions. What determines the number of rules to be tested? When should one stop generating rules? What is the justification for specifying particular expressions instead of any other expressions? FOCL, FOIL and "Discovery" system use different stop criteria and different mechanisms to generate rules for testing. RDM system "Discovery" selects rules, which are "law-like" (simplest and most general) and consistent with measurement scales [Krantz et all, 1971, 1989, 1990] for a particular task. The algorithm stops generating new rules when the rules become too complex (i.e., statistically insignificant for the data) in spite of the possibly high accuracy of the rules when applied to training data. FOIL and FOCL are based on the information gain criterion.

A RDM system "Discovery" contains several extensions over other FOL algorithms. It permits various forms of background knowledge to be exploited. The goal of the RDM system "Discovery" is to create probabilistic rules in terms of the relations (predicates and literals) defined by a collection of examples and other forms of background knowledge. RDM system "Discovery" as well as FOCL has several advantages over FOIL:

- Limits the search space by using constraints;
- Improves the search of hypotheses by using background knowledge with predicates defined by a rule directly in addition to predicates defined by a collection of examples;
- Improves the search of hypotheses by accepting as input a partial, possibly incorrect rule that is an initial approximation of the predicate to be learned.

There are also advantages of RDM system "Discovery" over FOCL:

- Limits the search space by using the statistical significance of hypotheses;
- Limits the search space by using the strength of data types scales;

The advantages above represent a way of generalization used in "Discovery" system. Generalization is the critical issue in applying data-driven forecasting systems. The RDM system "Discovery" generalizes data through "law-like" logical probabilistic rules presented in first order logic. Theoretical advantages of RDM system "Discovery" generalization are presented in [Vityaev, 1976, 1983, 1992, Vityaev et al, 1993, 1995; Kovalerchuk, 1973, Zagoruiko, 1976, Samokhvalov, 1973]. This approach has some similarity with the hint approach [Abu-Mostafa, 1990]. The main source for hints in first-order logic rules is representative measurement theory [Krantz et al., 1971, 1989, and 1990]. Note that a class of general propositional and first-order logic rules, covered by RDM system "Discovery" is wider than a class of decision trees.

RDM system "Discovery" selects rules, which are simplest and consistent with measurement scales (data types) for a particular task. Initial rule/hypotheses generation for further selection is problem-dependent. In [Kovalerchuk, Vityaev, 2000, chapter 5], we presented a set of rules/hypotheses specifically generated as an initial set of hypotheses for financial time series. This set of hypotheses can serve as a catalogue of initial rules/hypotheses to be tested (learned) for stock market forecasts.

The steps of MMDR are described in Figure 1. The first step selects and/or generates a class of logical rules suitable for a particular task. The next step learns the particular first-order logic rules using available training data. Then the first-order logic rules on training data using Fisher statistical test [Kendall, Stuart, 1977; Cramer, 1998] are tested. After that statistically significant rules are selected and Occam's razor principle is applied: the simplest hypothesis (rule) that fits the data is preferred [Mitchell, 1997, p. 65]. The last step creates interval and threshold forecasts using selected logical rules: IF A(x,y,…,z) THEN B(x,y,…,z).

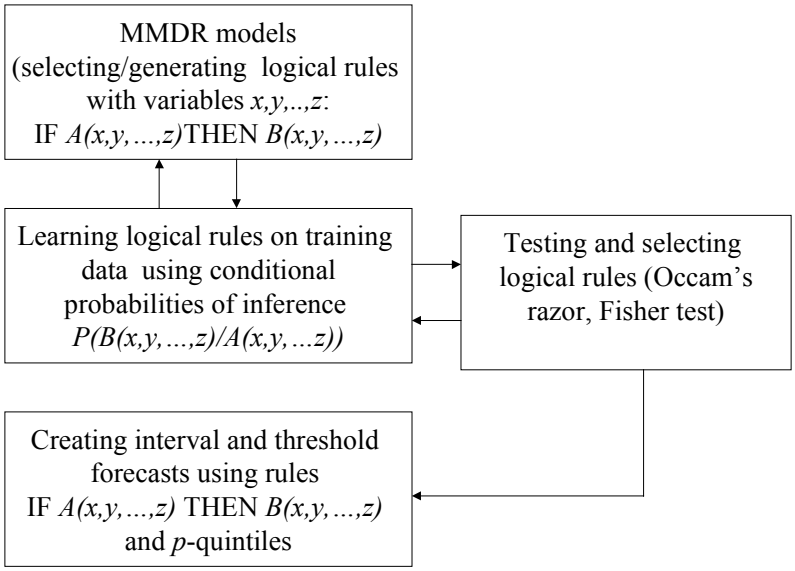

Figure 1. Flow diagram for MMDR: steps and technique

The original challenge for MMDR was the simulation of discovering scientific laws from empirical data in chemistry and physics. There is a well-know difference between "black box" models and fundamental models (laws) in modern physics. The latter have much longer life, wider scope, and a solid background. There is a reason to believe that MMDR caught some important features of discovering these regularities ("laws").

In this paper we consider the task of logically most powerful LET theory T discovery in the frame of the Relational Data Mining. This theory is the union of the "law-like" rules (see section 6), which are defined in this paper and satisfy all the properties of scientific laws: maximum generality, maximum refutability, simplicity and minimum number of parameters. The task was considered in conditions of the noise in data. For the most powerful theory discovery in the noise conditions the generalization of the "law-like" rules on probabilistic case is considered. Theory T provides the strongest predictions: it is proved (Vityaev, E.E., 1992) that it contains all assertions which predict some facts with maximum values of conditional probability.

5. Logical empirical theory

Let us introduce the first order logic L of the signature $\Im = \langle P_1,...,P_k\rangle$, $k > 0$, where $P_1,...,P_k$ are the predicate symbols of the arity $n_1...$, n_k. By an empirical system (Krantz, et al., v.1-3, 1971, 1989, 1990) we mean a finite model $M = \langle B,W\rangle$ of the signature $\Im$, where B - is the basic set of empirical system, $W = \langle P_1,...,P_k\rangle$ is the tuple of predicates of the signature $\Im$, defined on B. Each predicate P_j may be defined as a subset $P \subseteq B^{nj}$ on which it is true.

By an Logical Empirical Theory $T = \langle W, Obs, S^{\Im}\rangle$ we understand a triple consisting from a tuple of predicates W; measuring procedure Obs: $B \rightarrow \langle B^1,W\rangle$, mapping any finite subset of objects into the protocol of measurements represented by some finite empirical system $\langle B,W\rangle$; and $S^{\Im}$ – is the axiom system of the signature $\Im$, which must be true on any protocol of measurement. By the truth of some axiom on an empirical system M we mean a standard definition of the truth of expression on the model (empirical system).

The task of Logical Empirical Theory $T = \langle W, Obs, S^{\Im}\rangle$ discovery – is the task of the axiom system $S^{\Im}$ discovery on many-sorted empirical system $\mathcal{m} = \langle Å,W\rangle$. All results of observations Obs: $B \rightarrow \langle B,W\rangle$ are "parts" of that empirical system – any result of observation B is a finite submodel of the empirical system $\mathcal{m}$. If it is true, then it may be proved that the axiom system $S^{\Im}$ is universally quantifiable. Let $PR_{\mathcal{m}} = \{ Obs^w(B) \mid B \subset Å\}$ be the set of all experimental results that can be obtained as the protocols of observations on the finite sets of objects from empirical system $\mathcal{m}$.

Theorem 1 (Vityaev E., et al., 2003). If $PR_{\mathcal{m}}$ is the set of all finite submodels of an empirical system $\mathcal{m}$ then the axiom system $S^{\Im}$ is logically equivalent to the set of universal formulas.

We will consider only Logical Empirical Theories (LET) $\mathsf{T} = \langle W, Obs, S^{\mathfrak{I}} \rangle$ where the axiom system $S^{\mathfrak{I}}$ is universally quantifiable.

6. What is the law?

It may be proved that by logically equivalent transformations the set of universal formulas $S^{\mathfrak{I}}$ can be reduced to the set of rules of the form:

$$C = (A_1 \& ... \& A_k \Rightarrow A_0) \qquad (1)$$

where $A_j = P_j(x^j_1,...,x^j_{nj})^{\varepsilon j}$, $j = 0,1,...,k$; $x^0_1,...,x^0_{n0}$, $x^1_1,...,x^1_{n1}$,..., $x^k_1,...,x^k_{nk}$ - free variables; n0, n1,...,nk - arity of predicate symbols $P_1,...,P_k,P_0$; symbols $\varepsilon_0,\varepsilon_1,...,\varepsilon_k \in \{0,1\}$ means presence(0)/absence(1) of negation. We shall guess that the axiom system $S^{\mathfrak{I}}$ is a set of rules (1).

Thus, the task of Logical Empirical Theory T discovery is reduced to the task of the set of rules $S^{\mathfrak{I}}$ discovery. Let's analyse this task. What can we conclude about the truth of the axiom system $S^{\mathfrak{I}}$ on the set of experimental results $PR\mathcal{m}$, if guided only by the logical analysis of the axioms?

1. The rule $C = (A_1 \& ... \& A_k \Rightarrow A_0)$ can be true on $PR\mathcal{m}$ if the premise of the rule is always false on $PR\mathcal{m}$. We will prove that in that case some logically stronger "subrule" linking atoms of the premise is true on $PR\mathcal{m}$.
2. The rule C can be true on $PR\mathcal{m}$ because some of its logically stronger "subrule" containing only a part of the premise and the same conclusion is true on $PR\mathcal{m}$.

Let's clarify those logically more strong "subrules" from which the truth of the rule on $PR\mathcal{m}$ follows .

Theorem 2. The rule $C = (A_1 \& ... \& A_k \Rightarrow A_0)$ logically follows from any rule of the form:

(1) $A_{i1}\&...\&A_{ih} \Rightarrow \neg A_{i0}$,

where $\{A_{i1},...,A_{ih},A_{i0}\} \subset \{A_1,...,A_k\}$, $0 \le h < k$.

In that case we have $(A_{i1}\&...\&A_{ih} \Rightarrow \neg A_{i0}) \vdash \neg(A_1\&...\&A_k) \vdash (A_1\&...\&A_k \Rightarrow A_0)$;

(2) $(A_{i1}\&...\&A_{ih} \Rightarrow A_0)$,

where $\{A_{i1},...,A_{ih}\} \subset \{A_1,...,A_k\}$, $0 \le h < k$.

In that case we have $(A_{i1}\&...\&A_{ih} \Rightarrow A_0) \vdash (A_1\&...\&A_k \Rightarrow A_0)$.

where $\vdash$ is provability in a propositional calculus.

Definition 1. By *subrule* of some rule C we shall call any logically stronger rule of the form (1) or (2), defined in the theorem 2 for the rule C.

It is easy to see that any subrule also has the form (1).

Corollary 1. If some subrule of the rule C is true on PR_m, then the rule C is also true on PR_m.

Definition 2. By the *law* on the set of experimental results PR_m we shall call the rule C, which is true on PR_m, and for which none of its subrules is true on PR_m.

Let $\mathcal{L}$ be the set of all laws on PR_m. From the logic and methodology of science we know that only such hypotheses pretend to be laws that are the most falsifyable, simple and contain the least number of "parameters". In our case all these properties, which are usually difficult to define, follow from the logical strength of the laws. "Subrules" are simultaneously (1) logically stronger, than the rule and more falsifyable as they contain weaker premise and, therefore, are applicable to the greater volume of data; (2) more simple, as they contain smaller number of atomic expressions, than the rule; (3) include smaller number of

"parameters", as we may consider the extra number of atomic expressions as parameters of "tuning" the rule to data.

Why do we demand the laws should be most falsifiable, simple and contain the least number of parameters? Various authors uphold the different views on this subject. In our case for the hypotheses of the form (1) we can answer to this question. The discovery of laws allow us to solve more relevant task - to clarify, what is the most logically strong theory describing our data and probably lies in the bases of an unknown law of their generation.

Theorem 3. $\mathcal{L} \vdash S^{\Im}$.

Therefore initial task of LET T discovery reduced to the task of the most strong LET $T = \langle W, Obs^{v}, \mathcal{L} \rangle$ discovery.

7. Events and probability of events.

Let us generalize the notion of the law on probabilistic case. For that purpose let us introduce the probability on the set of experimental results and on logical expressions. Let us guess that the objects for the experiment are selected randomly from the set Å. It allows introduce the probability on the set of all experiments. For the case of simplicity following the work [7] we will introduce the discrete probability function on Å as a mapping μ: Å $\to$ [0,1] such that

$$\sum_{a \in \Lambda} \mu(a) = 1 \text{ and } \mu(a) \neq 0, \; a \in Å. \qquad (2)$$

For any B $\subseteq$ Å we will define $\mu(B) = \sum_{b \in B} \mu(b)$. Given the probability μ, we can thereby define a discrete probability function μ^n on the product of $(Å)^n$ by taking

$$\mu^n(a_1, ..., a_n) = \mu(a_1) \times ... \times \mu(a_n)$$

More general cases of the probability function μ definition are considered in [7].

Let's define the interpretation of the language L on empirical system $\mathcal{M} = \langle \text{Å}, W \rangle$ as a mapping I: $\Im \to W$, which associates with every signature symbol $P_j \in \Im$, j = 1,...,k, the predicate P_j from W of the same arity. Let $X = \{x_1, x_2, x_3, \ldots\}$ be the variables of the language L. By a valuation ν we mean a function $\nu: X \to \text{Å}$, mapping variables into the objects from Å.

Let's define a probability for sentences of the language L. Let $U(\Im)$ be the set of all atomic formulas of the language L of the form $P_j(x_1,\ldots,x_{nj})$; $\Re(\Im)$ is the set of all sentences of the language L, obtained by the closure of the set $U(\Im)$ relative to the logical operations &,∨,¬. By $\nu I\varphi$, $\varphi \in \Re(\Im)$ we will define the formula φ where the predicate symbols from $\Im$ replaced by the predicates from W by interpretation I and variables of the formula φ replaced by objects from Å by validation ν. Let us define the probability η of sentences from $\Re(\Im)$ for the empirical system $\mathcal{M}$. If $x_1,\ldots,x_n$ – all variables of the sentence $\varphi \in \Re(\Im)$, then

$$\eta(\varphi) = \mu^n(\{(a_1,\ldots,a_n) \mid \mathcal{M} \vDash \nu I\varphi, \nu(x_1) = a_1, \ldots, \nu(x_n) = a_n\}) \qquad (3)$$

8. General notion of law, probabilistic laws on $PR_{\mathcal{M}}$

Now we reformulate the concept of law on $PR_{\mathcal{M}}$ in terms of probability. Let's perform it in such a way that the concept of the law on $PR_{\mathcal{M}}$ would be a particular case of this definition. The law on $PR_{\mathcal{M}}$ is the rule, which is true on $PR_{\mathcal{M}}$ and all its subrules are false on $PR_{\mathcal{M}}$. Let us reformulate the concept of the law on $PR_{\mathcal{M}}$. The law is such rule true on $PR_{\mathcal{M}}$, which cannot be made simpler or logically more strength with saving the truth. This property of the law "to be not simplified" allows formulate the law not only in terms of a truth but also in terms of probability.

Theorem 4. Following two conditions are equivalent for any rule $C = (A_1\&...\&A_k \Rightarrow A_0)$:

1. the rule C is the law on PR$\mathcal{m}$;

2. (a) $\eta(A_0/A_1\&...\&A_k) = 1$ and $\eta(A_1\&...\&A_k) > 0$;

   (b) conditional probability $\eta(A_0/A_1\&...\&A_k)$ of the rule is strictly more than conditional probabilities of each of its subrules.

This theorem gives us equivalent definition of the law on PR$\mathcal{m}$ in the terms of probabilities.

Definition 3. By probabilistic law on PR$\mathcal{m}$ with conditional probability 1 we designate a rule $C = (A_1\&...\&A_k \Rightarrow A_0)$ satisfying following conditions:

a) $\eta(A_0/A_1\&...\&A_k) = 1$ и $\eta(A_1\&...\&A_k) > 0$;

b) conditional probability $\eta(A_0/A_1\&...\&A_k)$ of the rule is strictly more than conditional probabilities of each of its subrules.

Corollary 2. Probabilistic law on PR$\mathcal{m}$ with conditional probability 1 is the law on PR$\mathcal{m}$.

The task of the LET T discovery reduced therefore to the task of all probabilistic laws discovery with conditional probability equal to 1.

Definition of probability (3) base on the random choice of objects for the experiment. Experiment are “parts” (submodels) of the empirical system $\mathcal{M} = \langle \text{Å}, W\rangle$ and it is not supposed for them, that the truth values of predicates may be distorting during experiments as it takes place in the real experiments. The more general cases of the probability function μ definition, which considers the case of possible distortions of the experimental results presented in [7]. For such experiments “with noise” we can't demand the truth of the laws on PR$\mathcal{m}$. Therefore definition of the law on PR$\mathcal{m}$ loses the sense. The equivalent definition of the probabilistic

laws with conditional probability equal to 1 also loses sense - the conditional probability may be not equal to 1.

Let us consider the points 1 and 2 of the theorem 4 from the point of view of property of the law "not to be simplified":

- The law is such a rule, which is true on PR$\mathcal{m}$, which cannot be simplified or to be made logically more strongly without loosing the truth value.
- Any logically more strong subrule of the rule has strictly less conditional probability (less then 1), so the rule can't be simplified without the loosing of the value of conditional probability.

It gives us possibility to generalise the definition of the law:

Definition 4. Law is such a rule of the form (1), which has some estimation (truth, conditional probability, etc.) which we can't make logically more strong without loosing in the value of this estimation.

Therefore we may define the probabilistic law in more general case by deleting a condition (a) of definition 3 because it can't be satisfied in more general case. But we remain the condition (b), which express the property of the law in sense of more general definition 4.

Definition 5. By a probabilistic law on PR$\mathcal{m}$ we designate a rule $C = (A_1\&...\&A_k \Rightarrow A_0)$ of the form (1), which satisfy the condition: Conditional probability $\eta(A_0/A_1\&...\&A_k)$, $\eta(A_1\&...\&A_k) > 0$ of the rule is strictly more than conditional probabilities of each of its subrules.

Let's consider the task of the LET T discovery in presence of noise. This task was reduced to the task of all probabilistic laws with conditional probability 1 discovery. Is it possible to demonstrate the similar theorems in the case of noise presence? Generally it is impossible in

view of huge variety of possible models of noise, errors of probationer, models of instruments etc. Nevertheless our experience in practical tasks solving demonstrate that the concept of probabilistic laws is stable relative to some types of noise.

Definition 6. The noise model is said to be saving, if the set of probabilistic laws with probability 1 and the set of probabilistic laws are identical for this noise.

Therefore this work states the problem: to determine the set of saving models of noise. We found two examples of saving noise (Vityaev E., et al., 2003).

The task of LET T discovery in the presence of noise reduced thus to two tasks (1) determine that the noise is saving (2) to discover the set $\mathfrak{L}$ of all probabilistic laws. It follows from the theorem 3, theorem 4, corollary 2 and definition 6 that for saving noise we may completely discover the LET T, that is $\mathfrak{L} \vdash S^{\mathfrak{I}}$.

9. RDM "Discovery" system

A RDM system “Discovery” (Vityaev, 1976; Vityaev and Moskvitin, 1993) discovers probabilistic laws on data presented as many-sorted empirical system. This system has been successfully applied to many problems in medicine (cancer diagnostic systems), time series forecasting, psychophysics, and other fields (Kovalerchuk et al., 1996, 1997; Kovalerchuk and Vityaev, 2000; 2001; www-site "Scientific Discovery").

The “Discovery” system searches all chains C_1 , C_2 , …, C_{m-1}, C_m of probabilistic laws (see figure 2), such that:

1. each rule C_i is a subrule of the rule C_{i+1} ($C_i = sub(C_{i+1})$) as it is defined in theorem 2 point (2) see figure 2;
2. $Prob(C_1) < Prob(C_2), \dots , Prob(C_{m-1}) < Prob(C_m)$.

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. If there is no noise or we have a saving noise, all probabilistic laws with conditional probability 1 may be found at the end of such chains and LET T may be discovered. Otherwise the best approximation of the theory may be found (in the sense of the set rules which have maximum values of conditional probability).

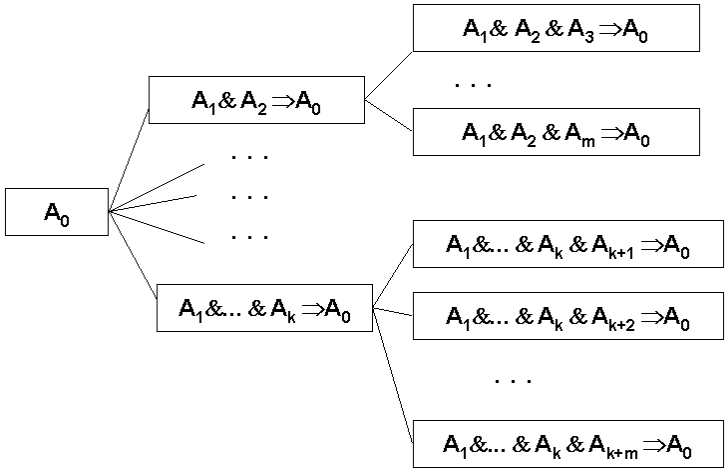

Figure 2. An example of the rule search for hypothesis A_0.

The algorithm stops generating new rules when they become too complex (i.e., statistically insignificant for the data). The Fisher statistical criterion is used in this algorithm for testing statistical significance.

ACKNOWLEDGEMENTS

This work was supported partly by grants from the Russian Foundation for Basic Research (grant #02-07-90355), NATO (grant LST.CLG 979815) and the Siberian Branch of the Russian Academy of Sciences (Integration project No. 119).

References

1. Abu-Mostafa, Y.S. (1990), `Learning from hints in neural networks', *J Complexity* 6, pp.192-198.
2. Bergadano, F., Giordana, A., & Ponsero, S. (1989). Deduction in top-down inductive learning. Proceedings of the Sixth International Workshop on Machine Learning (pp. 23--25). Ithaca, NY: Morgan Kaufmann.

3. Bratko, I., Muggleton, S., Varvsek, A. Learning qualitative models of dynamic systems. In Inductive Logic Programming, S. Muggleton, Ed. Academic Press, London, 1992
4. Bratko, I. Innovative design as learning from examples. In Proceedings of the International Conference on Design to Manufacture in Modern Industries, Bled, Slovenia, June 1993.
5. Bratko I, Muggleton S (1995): Applications of inductive logic programming. Communications of ACM 38 (11):65-70.
6. Carnap, R., Logical foundations of probability, Chicago, University of Chicago Press, 1962.
7. Chang C.C., Keisler H.J. Model theory. – Amsterdam, North-Holland, 1973.
8. Cramer D. (1998). Fundamental Statistics for Social Research, Step-by-step calculations and computer technique using SPSS for Windows, Routledge, London, NY
9. Danyluk, A. (1989). Finding new rules for incomplete theories: Explicit biases for induction with contextual information. Proceedings of the Sixth International Workshop on Machine Learning (pp. 34--36). Ithaca, NY: Morgan Kaufmann.
10. Dzeroski, S., DeHaspe, L., Ruck, B.M., and Walley, W.J. (1994). Classification of river water quality data using machine learning. In: Proceedings of the Fifth International Conference on the Development and Application of Computer Techniques to Environmental Studies (ENVIROSOFT'94).
11. Dzeroski S (1996): Inductive Logic Programming and Knowledge Discovery in Databases. In: Advances in Knowledge Discovery and Data Mining, Eds. U. Fayad, G., Piatetsky-Shapiro, P. Smyth, R. Uthurusamy. AAAI Press, The MIT Press, pp. 117-152.
12. Fenstad, J.I. Representation of probabilities defined on first order languages // J.N.Crossley, ed., Sets, Models and Recursion Theory: Proceedings of the Summer School in Mathematical Logic and Tenth Logic Colloguium (1967) 156-172.
13. Flann, N., & Dietterich, T. (1989). A study of explanation-based methods for inductive learning. Machine Learning, 4, 187--226.
14. Flach, P., Giraud-Carrier C., and Lloyd J.W. (1998). Strongly Typed Inductive Concept Learning. In Proceedings of the Eighth International Conference on Inductive Logic Programming (ILP'98), 185-194.
15. Fu LiMin (1999): Knowledge Discovery Based on Neural Networks, Communications of ACM, vol. 42, N11, pp. 47-50. Goodrich, J.A., Cutler, G., Tjian, R. (1996), `Contacts in context: promoter specificity and macromolecular interactions in transcription', Cell 84(6), 825-830.
16. Halpern, J.Y. (1990), `An analysis of first-order logic of probability', Artificial Intelligence 46, pp.311-350.

17. Hirsh, H. (1989). Combining empirical and analytical learning with version spaces.
18. Hyafil L, Rivest RL (1976): Constructing optimal binary decision trees is NP-Complete. Information Processing Letters 5 (1):15-17.
19. Katz, B.(1989). Integrating learning in a neural network. Proceedings of the Sixth international Workshop on Machine Learning (pp. 69--71). Ithaca, NY: Morgan Kaufmann.
20. Kendall M.G., Stuart A. (1977) The advanced theory of statistics, 4th ed., v.1.Charles Griffin & Co LTD, London.
21. Kovalerchuk B (1973): Classification invariant to coding of objects. Comp. Syst. 55:90-97, Institute of Mathematics, Novosibirsk. (in Russian).
22. Kovalerchuk, B., Talianski, V. (1992), `Comparison of empirical and computed fuzzy values of conjunction', Fuzzy Sets and Systems 46, pp.49-53.
23. Kovalerchuk, B., Triantaphyllou, E., Ruiz, J. (1996), `Monotonicity and logical analysis of data: A mechanism for evaluation of mammographic and clinical data' in R.F. Kilcoyne, J.L. Lear, A.H. Rowberg, eds., Computer Applications to assist Radiology, Carlsbad, CA: Symposia Foundation, pp.191-196.
24. Kovalerchuk, B., Vityaev, E., Ruiz, J.F. (1997), `Design of consistent system for radiologists to support breast cancer diagnosis' In: Proc Joint Conf Information Sciences, Durham, NC, 2, pp.118-121.
25. Kovalerchuk, B., Vityaev, E. (2000), Data Mining in finance: Advances in Relational and Hybrid Methods, Kluwer Academic Publishers, 308 p.
26. Kovalerchuk B., Vityaev E., Ruiz J. (2000), `Consistent Knowledge Discovery in Medical Diagnosis', IEEE Engineering in Medicine and Biology Magazine. Special issue: "Medical Data Mining", July/August 2000, pp.26-37.
27. Kovalerchuk, B., Vityaev, E., Ruiz, J.F. (2001), 'Consistent and Complete Data and "Expert" Mining in Medicine'. In: Medical Data Mining and Knowledge Discovery, Springer, pp.238-280.
28. Krantz, D.H., Luce, R.D., Suppes, P., Tversky, A. (1971, 1989, 1990), Foundations of measurement, Vol. 1,2,3 - NY, London: Acad. press, (1971) 577 p., (1989) 493 p., (1990) 356 p.
29. Lebowitz, M. (1986). Integrated learning: Controlling explanation. Cognitive Science, 10.
30. Mitchell, T. (1997), Machine Learning, New York: McGraw Hill.
31. Mooney, R., & Ourston, D. (1989). Induction over the unexplained: Integrated learning of concepts with both explainable and conventional aspects. Proceedings of the Sixth International Workshop on Machine Learning (pp. 5--7). Ithaca, NY: Morgan Kaufmann.

32. Muggleton S. (1994). Bayesian inductive logic programming. In Proceedings of the Eleventh International Conference on Machine Learning W. Cohen and H. Hirsh, Eds., pp. 371–379.
33. Muggleton S. (1999): Scientific Knowledge Discovery Using Inductive Logic Programming, Communications of ACM, vol. 42, N11, pp. 43-46.
34. Muggleton, S., & Buntine, W. (1988). Machine invention of first-order predicates by inverting resolution. Proceedings of the Fifth International Workshop on Machine Learning (pp. 339--352). Ann Arbor, MI: Morgan Kaufmann.
35. Muggleton, S., King, R.D. and Sternberg, M.J.E. (1992). Protein secondary structure prediction using logic. Prot. Eng. 5, 7), 647–657
36. Narens L. (1985), Abstract Measurement Theory, MIT Press, Cambridge.
37. Pazzani, M. (1989). Explanation-based learning with weak domain theories. Proceedings of the Sixth International Workshop on Machine Learning (pp. 72-- 74). Ithaca, NY: Morgan Kaufmann.
38. Pazzani, M. J. (1990). Creating a memory of causal relationships: An integration of empirical and explanation-based learning methods. Hillsdale, NJ: Lawrence Erlbaum Associates.
39. Pazzani, M., Brunk, C. (1990), Detecting and correcting errors in rule-based expert systems: An integration of empirical and explanation-based learning. Proceedings of the Workshop on Knowledge Acquisition for Knowledge-Based System. Banff, Canada.
40. Pazzani, M., Kibler, D. (1992). The utility of prior knowledge in inductive learning. Machine Learning, 9, 54-97
41. Pazzani, M., (1997), Comprehensible Knowledge Discovery: Gaining Insight from Data. First Federal Data Mining Conference and Exposition, pp. 73-82. Washington, DC
42. Pfanzagl J. (1971). Theory of measurement (in cooperation with V.Baumann, H.Huber) 2nd ed. Physica-Verlag.
43. Quinlan, J. R. (1989). Learning relations: Comparison of a symbolic and a connectionist approach (Technical Report). Sydney, Australia: University of Sidney.
44. Quinlan, J. R. (1990). Learning logical definitions from relations. Machine Learning, 5, 239-266.
45. Russel, S., Norvig, P. (1995), Artificial Intelligence. A Modern Approach, Englewood Cliffs, NJ: Prentice Hall.
46. Samokhvalov, K., (1973). On theory of empirical prediction, (Comp. Syst., #55), 3-35. (In Russian)

47. Sarrett, W., Pazzani, M. (1989). One-sided algorithms for integrating empirical and explanation-based learning. Proceedings of the Sixth International Workshop on Machine Learning (pp. 26--28). Ithaca, NY: Morgan Kaufmann.
48. Scott, D., Suppes P., (1958), Foundation aspects of theories of measurement, Journal of Symbolic Logic, v.23, pp. 113-128.
49. Shavlik, J., & Towell, G. (1989). Combining explanation-based learning and artificial neural networks. Proceedings of the Sixth International Workshop on Machine Learning ,pp. 90-93. Ithaca, NY: Morgan Kaufmann.
50. Suppes P., Zines J. (1963). Basic measurement theory. In: Luce, R., Bush, R., and Galanter (Eds). Handbook of mathematical psychology, v. 1, NY, Wiley, 1-76.
51. Thagard, P. & Shelley, C. (1997) Abductive reasoning: Logic, visual thinking, and coherence. In: M.-L. Dalla Chiara et al (eds), Logic and Scientific methods. Dordrecht: Kluwer, p.413-427. © Paul Thagard and Cameron Shelley, 1997
52. Vityaev, E.E. (1976), `Method of regularities determination and method of prediction', In: Empirical Prediction and Pattern Recognition Computational Systems, 67, pp.54-68. (in Russian).
53. Vityaev E. (1983). Data Analysis in the languages of empirical systems. Ph.D. Diss, Institute of Mathematics SD RAS, Novosibirsk, p.192. (In Russian)
54. Vityaev, E.E. (1992), `Semantic approach to knowledge base development: Semantic probabilistic inference', Computer Systems 146, pp.19-49. (in Russian).
55. Vityaev, E.E., Moskvitin, A.A. (1993), `Introduction to discovery theory: Discovery software system', Computational Systems 148, pp.117-163. (in Russian).
56. Vityaev E., Logvinenko A. (1995). Axiom systems testing method, Computational Systems, Theory of computation and languages of specification, (Comp. Syst., #152), Novosibirsk, p.119-139. (in Russian).
57. Vityaev E., Demenkov P. (2003). Empirical Theory Discovery. In: Probabilistic ideas in science and philosophy (Proceedings of the reagion conference, Novosibirsk, 23-26 sept., 2003), Novosibirsk, pp.86-89.
58. Zagoruiko N.G., Elkina V.N. Eds. (1976), Machine Methods for Discovering Regularities. Proceedings of MOZ'76, Novosibirsk. (In Russian)
59. Zhang, M.Q. (1998), `Identification of human gene core-promoters in silico', Genome Res 8, pp.319-326.
60. WWW-site "Scientific Discovery" http://www.math.nsc.ru/LBRT/logic/vityaev