Relational Methodology for Data Mining and Knowledge Discovery

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

[1] Sobolev Institute of Mathematics SB RAS, Novosibirsk State University, Novosibirsk, 630090, Russia.

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

Abstract.

Knowledge discovery and data mining methods have been successful in many domains. However, their abilities to build or discover a domain theory remain unclear. This is largely due to the fact that many fundamental KDD&DM methodological questions are still unexplored such as (1) the nature of the information contained in input data relative to the domain theory, and (2) the nature of the knowledge that these methods discover. The goal of this paper is to clarify methodological questions of KDD&DM methods. This is done by using the concept of Relational Data Mining (RDM), representative measurement theory, an ontology of a subject domain, a many-sorted empirical system (algebraic structure in the first-order logic), and an ontology of a KDD&DM method. The paper concludes with a review of our RDM approach and 'Discovery' system built on this methodology that can analyze any hypotheses represented in the first-order logic and use any input by representing it in many-sorted empirical system.

1. Introduction.

A variety of KDD&DM methods have been developed, however their abilities to build or discover a domain theory remain unclear. The goal of this paper is to clarify this and associated questions. We start from the questions about the nature of the information contained in input data relative to the domain theory.

Question 1. What is the nature of the information contained in input data relative to the domain theory and what is the empirical content of input data?

At first we note, that quantities in data are not numbers themselves, but numbers with an *interpretation*. For example, abstract numbers 200, 3400, 300, 500 have three different interpretations shown in Table 1.

Table 1.

Interpretation	*Values*	*Meaningful operations*
Abstract numbers	200, 3400, 300, 500	Meaning of 20 > 30 is not clear. These numbers can be just labels.
Abstract angles	200, 3400, 300, 500	300 meaningfully greater that 200
Azimuth angles	200, 3400, 300, 500	Azimuth operations
Rotational angles	200, 3400, 300, 500	Rotational angle operations

For every quantity there are relations and operations that are meaningful for this quantity. This interpretation of quantities is a core approach of the Representational Measurement Theory (RMT) [26], [37]. The RMT interprets quantities as *empirical systems* – algebraic structures defined on the objects of subject domain with the set of empirically interpretable relations and operations that define the meaning of the quantity (see the next section for definitions). The empirical content of quantities, laws and data is expressed by their (many-sorted) empirical systems.

[1] Corresponding author. E-mail: vityaev@math.nsc.ru

More specifically main statements of the measurement theory relative to data mining issues are as follows [26], [37]:

- numerical representations (scales) of quantities, laws and data are determined by the corresponding empirical systems;
- scales are unique up to a certain sets of permissible transformations such as changing measurement units from meters to kilometers for ratio-scales;
- the laws and KDD&DM methods need to be invariant relative to the sets of permissible transformations of quantities in data.

To represent the empirical content of data in accordance with the measurement theory we need to transform the data into many-sorted empirical systems. These transformations are described in [23] for such data types as pair comparisons, binary matrices, matrices of orderings, matrices of proximity and attribute-based matrix.

The problem of extraction the empirical content of data and transformation of it into the many-sorted empirical systems for the general case is considered in section 3. This transformation depends on the subject domain. For example, for many physical quantities, there exists the interpretable operation $\bullet$, which possess all formal properties of additive operation +. However, medicine and other areas may have no empirical interpretation for the operation $\bullet$. For example, empirical system of temperature, measured by a thermometer, may include operation $\bullet$ that produces it temperature t_3 from temperatures t_1 and t_2, such that $t_3 = t_1 \bullet t_2$. But this operation in medicine may be non interpretable. In that case, the $\bullet$ operation should be removed from the empirical system of temperature in medicine and the corresponding scale should be reconsidered. The order relation for temperature is obviously interpretable in medicine, and the order relation may be included in the empirical system of temperature in medicine. Physical temperature measured by thermometer for the case of medicine is for example the indirect measure of the metabolic rate of the patient. Thus, the empirical content of data and scales depend on the interpretation.

Question 2. What determines the interpretation of data and what information may be extracted from data?

The interpretation depends on ontology, which determines the view of the 'real-world'. The subject domain and ontology are closely connected. The patient in medicine may be considered from many points of view: psychology, physiology, anatomy, sociology and so on. Thus, we cannot properly extract information from data about the patient without setting up an ontology.

Consider another example from the area of finance. What is the empirical content of financial time series? It also can be viewed from many points of view that include the points of view of (1) a trader-expert, (2) one of the mathematical disciplines, (3) technical analysis, (4) trade indexes etc. We need to specify the ontology to extract the empirical content and interpret all relations and operations to define the meaning of all quantities contained in data. The information extracted from data is represented by many-sorted empirical system with that set of relations and operations.

Question 3. What is the nature of the knowledge that a particular KDD&DM method extracts from input data?

As was pointed out above any KDD&DM method contains a view of the "real-world" (ontology of the method). More completely, from our viewpoint, any KDD&DM method explicitly or implicitly assume:

(1) some quantities for input data;
(2) ontology of particular KDD&DM method (including a language) to manipulate and interpret data and results;
(3) class of hypothesis in terms of that language to be tested on data (knowledge space of the KDD&DM method, see definition in section 5);

(4) Interpretation of the ontology of particular KDD&DM method in the ontology of the subject domain. For example, to apply a classification method that uses spherical shapes for classes of data, we need first to interpret spherical shapes in the ontology of the subject domain. Otherwise, we cannot interpret the results of classification.

The knowledge extracted by a KDD&DM method is the set of confirmed hypothesis that are interpretable in the ontology of the KDD&DM method and in the ontology of the subject domain.

The results of the KDD&DM methods should not depend on the subjective choice of the measurements units of quantities contained in data. Thus, KDD&DM methods need to be invariant relative to a sets of permissible transformations of quantities. In section 4, we define the invariance of KDD&DM methods relative to permissible transformations of scales. But, as we pointed out above, the scales of quantities depend on the interpretation of the relations and operations with data. Hence, the invariance of the method cannot be established before we revise the scales of all quantities based on the interpretation of their relations and operations. For example, we need to determine a new scale for temperature relevant to medical data before we can establish the invariance of the KDD&DM method for such data. We discuss this issue in section 3.

The invariance of a KDD&DM method is closely related to interpretability of its results in the ontology of the subject domain. If a KDD&DM method uses operations or relations, that are not interpretable in the language of the method (and hence in the subject domain ontology), then it may obtain non-interpretable results. For example, average or sum of patients' temperatures in the hospital has no medical interpretation. If all relations and operations, used in the algorithm, are included in the empirical systems of quantities, then the algorithm will be obviously invariant relative to the permissible transformations of scales for these quantities.

Thus, to avoid the non invariance of the method and non interpretability of its results we need to use in the algorithms only relations and operations that are interpretable from the many-sorted empirical system. It means that the hypotheses tested by the method include only those interpretable in the ontology of the subject domain relations and operations. Then the knowledge space of particular KDD&DM method is a set of hypothesis formulated in terms of the relations and operations interpretable in the ontology. According to the measurement theory, any numerical data type can be transformed into a relational form that preserves all relevant information of that numerical data type.

Now we can formulate our answer to the question about the nature of the knowledge that KDD&DM methods extract -- *any KDD&DM method extracts a confirmed hypothesis from its knowledge space*. In this context, the main characteristics of each KDD&DM method are *the method's ontology and knowledge space.*

Discoveries of many KDD&DM methods are not invariant to the permissible transformations of scales of input data. As a result, these discoveries are not fully interpretable in the subject domain ontology. The interpretation usually is not explicitly defined and may be subjective. The expert in the subject domain can obtain a correct conclusion using non invariant method by using informal intuitive extraction of interpretable results from the confirmed hypotheses, but this is rather an art of Data Mining.

The Relational Data Mining (RDM) approach [24-25, 51-52] and "Discovery" system described below allow us to overcome presented limitations. This approach is based on the first-order logic for knowledge extraction from the many-sorted empirical systems for various classes of hypothesis.

The relational approach:

a) extending the data type notion;

b) using first-order logic and the measurement theory for presenting various data types as many-sorted empirical systems;
c) using any background knowledge expressed in the first-order logic for learning and forecasting;
d) extracting various hypotheses classes formulated in the first-order logic.

This approach allows us to answer to the following question.

Question 5. Can a subject domain theory be discovered by using KDD&DM methods?

As was pointed out above any KDD&DM method discovers some class of hypothesis. In the relational data mining approach, classes of hypotheses are extended to classes of hypotheses presented in the first-order logic. Several KDD&DM methods discover wide classes of hypotheses in the First-Order Logic (FOL). We compare some FOL methods with our relational approach and the "Discovery" system in section 6.

In sections 7-14, we present a series of definitions and prove that a subject domain theory can be discovered by using the relational data mining approach. We define notions of law and probabilistic law (section 10), a subject domain theory T as the set of all laws, and the theory TP as the set of all probabilistic laws. Next, a theorem is proved that the set SPL of Strongest Probabilistic Laws (with the maximum values of conditional probability) contains a subject domain theory T, and $T \subset SPL \subset TP$. We define a notion of a *most specific rule* (section 13) and prove that an inductive statistical inference (based on these rules) avoids the problem of statistical ambiguity. Next in sections 12 and 13, a special semantic probabilistic inferences is defined that infers theories T, TP, set SPL and generalizes the logical inference of logic programming.

Finally, we describe the "Discovery" system, which implements semantic probabilistic inferences and discovers theories T, TP and sets SPL, MSR. As a result, it is proved that subject domain theory T, its probabilistic approximation TP and its consistent probabilistic approximation MSR can be discovered in the frames of the relational approach using the "Discovery" system. The "Discovery" system has been successfully applied to solutions of many practical tasks (see website www.math.nsc.ru/AP/ScientificDiscovery).

The original challenge for the RDM "Discovery" system was the simulation of discovering scientific laws from empirical data in chemistry and physics. There is a well-known difference between the "black box" models and basic models (laws) in modern physics. The lifetime of the latter models is much longer, the scope is wider, and their background is sound. There is reason to believe that the RDM "Discovery" system captures certain important features of the discovery of laws.

2. Representative Measurement Theory

In accordance with the measurement theory, numerical representations of quantities, laws and data are determined by the corresponding empirical systems. In this section we present required definitions from the measurement theory [26],[37].

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

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

Every relation P_i is a Boolean function (a predicate) with k(i) arguments from A, and ρ_j is the k(j) argument operation on A. A system **R**

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

is called a *numerical system of the same type as system* **A,** if R is a subset of Re^m , m ≥ 1, Re^m is a set of m-tuples of real numbers, every relation T_i has the same arity k(i) as the corresponding relation P_i, and every real-value function σ_j has the same arity k(j) as the corresponding operation ρ_j.

A numerical system **R** is called a *numerical representation* of the empirical system **A,** if a (strong) homomorphism φ: A → R exists such that:

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

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

The strong homomorphism means that if predicate $T_i(\varphi(a_1),\ldots, \varphi(a_{k(i)}))$ is true on $\langle\varphi(a_1),\ldots, \varphi(a_{k(i)})\rangle$, then there exists tuple $\langle b_1,\ldots,b_{k(i)}\rangle$ in A, such that $P_i(b_1,\ldots,b_{k(i)})$ is true and $\varphi(b_1) = \varphi(a_1),\ldots, \varphi(b_{k(i)}) = \varphi(a_{k(i)})$. We will denote such homomorphism between the empirical system **A** and numerical system **R** as φ: **A** → **R**. Thus, the numerical system **R** represents a relational structure in computationally tractable form with a complete retention of all the properties of the relational structure.

In the measurement theory, the following process is in place:

(1) finding a numerical representation **R** for empirical system **A**;

(2) proving a theorem that homomorphism φ: **A** → **R** exists; and

(3) defining the set of all possible transformations f: **R** → **R** (the uniqueness theorems) of the homomorphism φ, such that fφ is also homomorphism fφ: **A** → **R**.

Example: A relational structure **A** = ⟨A, P⟩ is called a semi-ordering, if for all a, b, c ∈ A the following axioms are satisfied:

$(P(a,b)\&P(b, c) \Rightarrow \forall d \in A(P(a, d)\vee P(d, c)))$.

Theorem [10]: If **A** = ⟨A, P⟩ is semi-ordering, then there exists a function U: A → Re such that:

$P(a,b) \Leftrightarrow U(a) + 1 < U(b)$.

There are hundreds of numerical representations known in the measurement theory with few most commonly used. The strongest one is called the absolute data type (*absolute scale*). The weakest numerical data type is the nominal data type (*nominal scale*). There is a spectrum of data types between them. They allow us comparing, ordering, adding, multiplying, dividing values and so on. The classification of these data types is presented in table 1. The basis of this classification is a transformation group. The strongest absolute data type does not permit to transform data at all, and the weakest nominal data type permits any one-to-one transformation. Intermediate data types permit different transformations such as positive affine, linear and others (see table 1).

Table 1. Numerical data types.

Transformation	Transformation Group	Data type (scale)
$X \rightarrow f(x)$,	F :Re → (onto)Re, 1→1 transformation group	Nominal
$X \rightarrow f(x)$,	F :Re → (onto)Re homeomorphism group	Order
$X \rightarrow rx + s$, $r > 0$	Positive affine group	Interval
$X \rightarrow tx^r$, $t,r > 0$	Power group	Log-interval
$X \rightarrow x + s$	Translation group	Difference
$X \rightarrow tx$, $t > 0$	Similarity group	Ratio
$X \rightarrow x$	Identity group	Absolute

The transformation groups are used to determine the invariance of law. The law expression must be invariant to the transformation group; otherwise it will depend not only on the nature, but on the subjective choice of the of measurement units.

3. Data type problems

Below we consider a problem of the empirical content of data. A data type in the object-oriented programming languages as well as in the measurement theory is relational structure **A** with the sets of relations and operations interpretable in the domain theory. For instance, a "stock price" data type can be represented as a relational structure $\mathbf{A} = \langle A; \{\leq, =, \geq\}\rangle$ with where nodes A are individual stock prices and arcs are their relations $\{\leq, =, \geq\}$. Implicitly, every attribute in the data represents a data type, which can take a number of possible values. These values are elements of A. For instance, the attribute "date" has 365 (366) elements ranging from January 1 to December 31. There are several meaningful relations and operations with dates such as $<$, $=$, $>$, 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. Conventionally, in attribute-value languages (AVL), this data type as well as many other data types are given implicitly, i.e., the relations and operations are not explicitly presented.

Below we consider this implicit situation in more detail for six cases:

1. Physical data types in physical domains.
2. Physical data types in non-physical domains.
3. Non-physical data types in non-physical domains
4. Nominal discrete data types.
5. Non-quantitative and non-discrete data types.
6. Mix of data types.

1. Physical data types in physical domains. Data contain only physical quantities, ontology and physics domain background knowledge of the learning task. This is a realm of physics with well-developed data types and measurement procedures. In this case, the measurement theory [26] provides formalized relational structures for all the quantities and KDD&DM methods can be correctly applied.

2. Physical data types for non-physical domains. The data contain physical quantities, but the ontology and domain background knowledge of the learning task does not refer to physics. The background knowledge may refer to finance, geology, medicine, and other areas. In such cases, the real data types are not known, even when they represent physical quantities (as we pointed out above for the temperature in medicine). If the quantity is physical, then we can define the relational structure from structures available in the measurement theory. However, the physically interpretable relations of the relational structure are not necessarily interpretable in ontologies of other subject domains. Interpretation of the relations and operations should be provided for a new domain. If relations are not interpretable, they should be removed from the relational structure. The invariance of the KDD&DM results is not guarantied if the relation is not removed and a data mining method uses it (see the next section for definitions).

3. Non-physical data types in non-physical domains. For non-physical quantities, data types are virtually unknown. There are two sub-cases:

a. Non-numerical data types. It has been demonstrated in [23] that several data types such as pair-wise and multiple comparison data types, attribute-based data types, order, and coherence matrixes data types can be represented in many-sorted empirical systems in the rather natural way. Without such representation, the invariance of KDD&DM methods cannot be rigorously established.

b. Numerical data types. Here, we have a measurer x(a), which produces a number as a result of a measurement procedure applied to an object *a*. Examples of measurers are psychological tests, stock market indicators, questionnaires, and physical measuring instruments 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 $R(y_1,\ldots,y_k) \subset Re^k$ and operation $\sigma(x_1,\ldots,x_m): Re^m \to Re$, where Re is the set of real numbers, an *empirical relation* P^R on A^k and an *empirical operation* $\rho^\sigma: A^m \to A$ can be defined as follows

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

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

The values x(a) provided by the measurer obviously have an empirical interpretation, but the relation P^R and operation ρ^σ may not. We should find relations R and operations σ that have an empirical interpretation in the subject domain ontology. The set of derived interpretable relations is not empty, because at least one relation ($P^=$) has an empirical interpretation: $P^=(a_1,a_2) \Leftrightarrow x(a_1) = x(a_2)$.

In the measurement theory, many sets of axioms that establish strong data types are based only on ordering and equivalence relations. Some strong data types can be constructed from interactions of the quantities with weak data types, such as ordering and equivalence.

For instance, given weak order relation $<_y$ (for the attribute y) and n equivalence relations $\approx_{x_1}, \ldots, \approx_{x_n}$ for the attributes $x_1,\ldots,x_n$, one can construct a complex relation $G(y,x_1, \ldots,x_n) \Leftrightarrow y = f(x_1, \ldots, x_n)$ (defined by the axiomatic system) between y and $x_1,\ldots,x_n$, such that $f(x_1,\ldots,x_n)$ is a polynomial [26]. For the polynomial the multiplication, power and sum operations are required. However, these operations can be defined for y, $x_1,\ldots,x_n$ using relation G, if a certain set of axioms in terms of relations $<_y$, $\approx_{x_1}, \ldots, \approx_{x_n}$ is true for A.

Ordering and equivalence relations are usually empirically interpretable in the ontology of various subject domains. The invariance of the KDD&DM is not guarantied for initial numerical data types for which they usually applied, but is guarantied for revised scales, based on relations P^R and operations ρ^σ.

4. Nominal discrete data types. Here, data are interpretable in the corresponding relational structures, because there is no difference between the numerical and empirical systems beyond possible use of different symbols. All numbers can be considered as names, and can be easily represented as predicates with a single variable. The KDD&DM methods and their results are invariant if discrete data types are used as names.

5. Non-quantitative and non-discrete data types. Data contain no quantities and discrete variables, but do contain ranks, orders and other non-numerical data types. This case is similar to the above item 3a. The only difference is that such data are usually made discrete by various calibrations with a loss of useful information.

6. Mix of data types. All the mentioned difficulties arise in this case. To work with such mix requires a new approach. Our Relational Data Mining approach provides it using a relational representation of the data types.

4. Invariance of the KDD&DM methods

The results of the KDD&DM methods must not depend on the subjective choice of the measurement units, but usually it is not the case. Let us define the notion of invariance of a KDD&DM method. To that end, we will use the common (attribute-based) representation of a supervised learning [54] (fig. 1), where:

- W={w} is a training sample;
- $X(w) = (x_1,\ldots,x_n)$ is the tuple of values of n variables (attributes) for training example w;
- Y(w) is the target function assigning the target value for each training example w;

The result of KDD&DM method M learning on the training set {X(w)}, $w \in W$ is a rule J

$$J = M(\{X(w)\}),$$

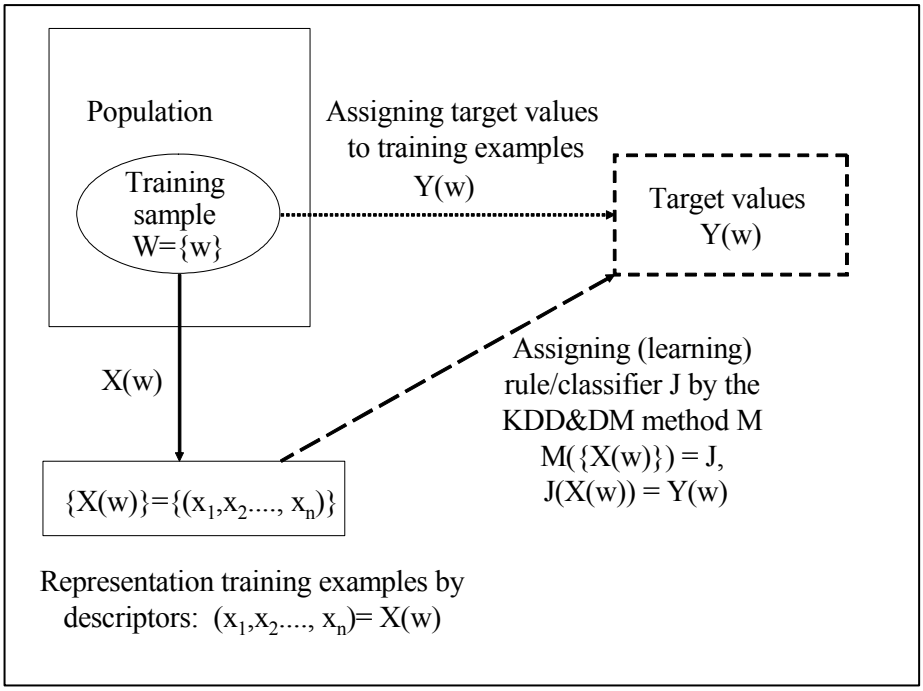

Figure 1. Schematic supervised attribute-based data mining model

that predicts values of the target function Y(w). For example, consider w with unknown value Y(w) but with the values all its attributes X(w) known, then

$$J(X(w)) = Y(w),$$

where J(X(w)) is value generated by the rule J. The resulting rule J can be an algebraic expression, a logic expression, a decision tree, a neural network, a complex algorithm, or a combination of these models.

If the attributes $(x_1,\ldots,x_n, Y)$ are determined by the empirical systems $A_1, \ldots, A_n$, and B having the transformation groups $g_1, \ldots, g_n$, and g, respectively, then the transformation group G for all attributes is a product $G = g_1\times\ldots\times g_n\times g$.

The KDD&DM method M is invariant relative to the transformation group G iff for any $g\in G$ rules

$$J = [M(\{X(w)\})],\ J_g = [M(\{g(X(w))\})],$$

produced by the method M, generate the same results and for any X(w), $w \in W$

$$gJ(X(w)) = J_g(g(X(w)).$$

If the method is not invariant (that is the case for majority of the methods), then predictions generated by the method depends on the subjective choice of the measurement units.

The invariance of the method is closely connected to the interpretability of its results.

The numerical KDD&DM methods assume that a numerical standard mathematical operation such as +,-,*, / can be used in an algorithm despite possible non-interpretability. In this case, the method can be non-invariant and can deliver non-interpretable results. In contrast a KDD&DM method M is invariant if it uses only information from empirical systems $A_1, \ldots, A_n$, and B as data and produces rules J that are logical expressions in terms of these empirical systems. This approach is proposed in Relational approach to Data Mining [23],[25], [47], [51].

5. Relational methodology for the analysis of KDD&DM methods

A non invariant KDD&DM method M: $\{X(w)\} \rightarrow J$ can be analyzed and another invariant method can be extracted from M. Let us define a many-sorted empirical system A(W) that is a product of empirical systems $A_1,\ldots,A_n$, B bound on the set W. The empirical system A(W) contains all interpretable information from a learning sample W. Next we define transformation $W \rightarrow A(W)$ of all data W into a many-sorted empirical system A(W), and replace the representation

$$W \rightarrow \{X(w)\}$$

by the transformation

$$W \rightarrow A(W) \rightarrow \{X(w)\}.$$

Based on method M: $\{X(w)\} \rightarrow J$, we define a new method ML: $A(W) \rightarrow J$ such that

$$ML(A(W)) = M(\{X(w)\}) = J,$$

using transformation $W \rightarrow A(W) \rightarrow \{X(w)\}$. Thus, method ML uses only interpretable information from data A(W) and produces the rule J using method M.

Let us analyze the transformation of the interpretable information A(W) into the rule J through the method M. If we apply only interpretable operations to the interpretable information A(W) from the method M, we may extract some logical rule JL from rule J. This rule JL will contain only interpretable information from the rule J, expressed in terms of empirical system A(W).

Let us define the next method

$$MLogic: A(W) \rightarrow JL,$$

where the rule JL is a set of logical rules (1) for rule J produced by method M, and interpretable information A(W),

$$A(W) \rightarrow \{X(w)\}, M: \{X(w)\} \rightarrow J.$$

The method MLogic is obviously invariant. If we consider all possible data for the method M, and all rules JL, that may be produced by the MLogic method, then we will obtain a class of rules (hypotheses) {JL} (**knowledge space**) of the KDD&DM method M. As a result we obtain: (1) an ontology of the particular KDD&DM method M as an empirical system A(W), and (2) a knowledge space {JL} of the method M.

There are First-Order Logic (FOL) methods that are capable discovering some sets of hypotheses (knowledge space) presented in the first-order logic. Thus, these FOL methods that are invariant simulate (model) traditional KDD&DM methods at some extent.

6. First-order logic approaches

Let us consider the existent FOL-methods and compare them with the proposed relational approach and the Discovery system.

A variety of relational KDD&DM systems has been developed in recent years [29]. Theoretically, they have many advantages. However, in practice, the complexity of the language is greatly restricted reducing its applicability. For example, some systems require that the concept definition be expressed in terms of attribute-value pairs [27][6] or only in terms of unary predicates [17],[30][19][43][41]. The systems that allow workable relational concept definitions (e.g., OCCAM [33][11][2]) place strong restrictions on the form of induction and the initial knowledge that is provided to the system [35].

The major successful applications of the FOL have been described in [3],[4],[7],[8], [21], [32],[36] that are in domains of chemistry, physics, medicine and others. Tasks, such as mesh design, mutagenicity, and river water quality, exemplify successful applications. Domain specialists appreciate that the learned regularities are understandable directly in domain terms. Fu [13] has justly observed: "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. Comprehensive predicate invention. Human-readable and comprehensible form of rules. Logical relations (predicates) should be developed to exploit advantages of the human-readable forecasting rules. In this way, FOL methods can produce valuable comprehensible rules in addition to the forecast. Using this technique, a user can evaluate the performance of a forecast as well as a forecasting rule. Obviously, comprehensive rules have advantages over a forecast without explanations. The problems of inventing predicates were considered in the previous section and in [23].

Advantages versus disadvantages of the attribute-value languages (AVLs) methods and the first-order logic methods (table 1). Bratko and Muggleton [4] have indicated that the current 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 the current FOL methods. There are two types of numerical data in data mining:

(a) a numerical target variable;

(b) numerical attributes used to describe objects and discover patterns.

Traditionally, the FOL methods solve only classification tasks without direct operations on the numerical data. The 'Discovery' system handles an interval forecast of continuous numerical variables, like prices, along with classification tasks. The 'Discovery' system handles numerical time series using the first-order logic techniques, which is not typical of ILP and FOL applications.

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

Method	*Advantages for the learning process*	*Disadvantages for the learning process*
Method based on attribute-value languages	**Simple**, **efficient**, and handles noisy data.	**Limited form** of background **knowledge**. **Lack of relations** in the concept description language.
Method based on the First Order Logic	Appropriate learning time with a high **number of training examples**. Sound **theoretical basis** (first-order logic, logic programming). **Flexible** form of background knowledge, problem representation, and problem-specific constraints. **Comprehensive representation** of background knowledge, and relations among examples.	Inappropriate learning time with a high **number of arguments** in the relations. **Poor facilities** for processing **numerical data without using the measurement theory.**

Background knowledge and ontology. Knowledge Representation is an important and informal first step in Relational Data Mining. In the attribute-based methods, the attribute form of data actually dictates the form of knowledge representation. Relational data mining has more options for this purpose. For example, for RDM the attribute-based stock market information, such as stock prices, indices, and volume of trading should be transformed into the first-order logic form. This knowledge includes much more than just attribute values. There are many ways to represent knowledge in the first-order logic language. Data mining algorithms may work too long to "dig out" relevant information or can even produce inappropriate rules. Introducing data types [12] and concepts of the representative measurement theory [26][37] into the knowledge representation process helps to address this representation problem. In fact, the measurement theory developed a wide set of data types, which cover the data

types used in [12]. The FOL systems have a mechanism to represent background knowledge in a human-readable and comprehensive form.

Hybridizing the logical data mining methods with a probabilistic approach. This is done by introducing probabilities over logical formulas [5][9] [14] [21] [23][25] [31][45] [47] [48]. For financial data mining this was done in [22][23] [45][47][48] using the 'Discovery' system that has been applied to predict the SP500C time series and to develop a trading strategy. This RDM method outperformed several other strategies in simulated trading [22][23].

Statistical significance. Traditionally, the FOL methods were purely deterministic, which originated from logic programming. The deterministic methods have a well-known problem of handling data with a significant level of noise. This is especially important in financial data, which are very noisy. In contrast, the RDM can handle noisy and imperfect data including numerical data. Statistical significance is another challenge for deterministic methods. Statistically significant rules are advantageous in comparison with rules tested only for their performance on training and testing data [29]. Training and testing data can be too limited and/or not representative. There are more chances that rules would fail to give a correct forecast on other data if we .rely only on them,. This is a difficult problem for any data mining method, especially for deterministic methods, including ILP. Intensive studies have been conducted for incorporating a probabilistic mechanism into the ILP [31].

Hypotheses space. It is well known that the general problem of rule generating and testing is NP-complete [18]. Therefore, the above discussion is closely related to the following questions. What determines the number of rules? When to stop generating rules?

The number of hypotheses is another important parameter. It has already been mentioned that the RDM with the first-order rules allows expressing naturally a wide variety of general hypotheses. These more general rules can be used in solving classification problems as well as for interval forecasting of continuous variables. The algorithmic complexity of FOL algorithms is growing exponentially with the number of combinations of predicates to be tested. A restraint to halt this exponential growth is required in order to reduce a set of combination. To address this issue, we propose an approach based on data types and the measurement theory. This approach provides better means for generating 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 single examples are not generalized beyond the limits of statistically significant rules.

FOIL, FOCL and 'Discovery' algorithms. The algorithm FOIL [38][39] learns constant-free Horn clauses, a useful subset of first-order predicate calculus. Subsequently, the FOIL was extended to use a variety of types of background knowledge to increase the class of problems solvable, to decrease the hypothesis space explored, and to improve the accuracy of learned rules.

The FOCL (First Order Combined Learner) algorithm [35] extends FOIL. FOCL uses the first-order logic and combines its information based optimality metric with background knowledge. The FOCL has been tested on various problems [36] that include a domain describing when a student loan is to be repaid [34].

As indicated above, the general problem of rule generating and testing is NP-complete. Therefore, we face the problem of designing NP-complete algorithms. Several related questions are raised. What determines the number of rules to be tested? When to stop generating rules? What is the justification for specifying particular expressions instead of any other? The FOCL, FOIL and 'Discovery' systems use different stop criteria and different mechanisms to generate rules for testing. The RDM 'Discovery' system selects rules, which are probabilistic laws (see section 10) and consistent with measurement scales [26] for a particular task. The algorithm stops generating new rules when they become too complicated (i.e., statistically

insignificant for the data) despite the possibly high accuracy of the rules when applied to training data. The FOIL and FOCL are based on the information gain criterion.

The 'Discovery' system contains several extensions over other FOL algorithms. It enables various forms of background knowledge to be exploited. The goal of this system is to create probabilistic laws in terms of the relations (predicates and literals) defined by a collection of examples and other forms of background knowledge.

The 'Discovery' system, as well as FOCL, have several advantages over FOIL:

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

There are also advantages of this RDM system over FOCL which.

- Limits the search space by using the statistical significance of hypotheses and
- Limits the search space by using the strength of the data type scales.

The above advantages are ways of generalization used in this system. Generalization is the critical issue in applying the data-driven forecasting systems. The Discovery system generalizes data through probabilistic laws (see below). The approach is somewhat similar to the hint approach [1]. The main source for hints in the first-order logic rules is the representative measurement theory [26]. Note, a class of general propositional and first-order logic rules, covered by the system is wider than the class of decision trees.

7. Subject domain theory

In this section, we introduce the notion of subject domain theory and define the property of an experiment, which necessitate the universal axiomatizability of this theory.

Let us introduce the first-order logic L of signature $\Im = \langle P_1,...,P_k\rangle$, $k > 0$, where $P_1,...,P_k$ are predicate symbols of the arity $n_1..., n_k$. An empirical system [37][26] is a finite model $M = \langle B, W\rangle$ of the signature $\Im$, where B is the basic set of the empirical system, $W = \langle P_1,...,P_k\rangle$ is the tuple of predicates of the signature $\Im$ defined on B. Every predicate P_j can be also defined as a subset $P \subseteq B^{nj}$ on which it is true.

Let us represent a subject domain by the many-sorted empirical system $M = \langle A, W\rangle$. A Subject Domain Theory (SDT) $T = \langle A, W, Obs, S^{\Im}\rangle$ is a set that consists of:

- the tuple of predicates W of the signature $\Im$;
- the measuring procedure Obs: $B \rightarrow \langle B,W\rangle$, mapping any finite subset of objects $B \subset A$ into the protocol of measurements, represented by a finite (many-sorted) empirical system (data) $D = \langle B,W\rangle$;
- axiom system $S^{\Im}$, that should be true on any protocol of measurement. The defintion of the truth of an axiom in an empirical system D is a standard definition of the truth of expression on a model (empirical system).

Task: a task of discovering a subject domain theory is determining a system of axioms $S^{\Im}$ that is true on data (presented by a many-sorted empirical system $D = \langle B, W\rangle$).

All observation results Obs: $B \rightarrow \langle B, W\rangle$ are "parts" of empirical systems – any observation result is a finite submodel of the subject domain M. In this case, we can prove that the axiomatic system $S^{\Im}$ is universally axiomatizable. Let $PR_M = \{Obs(B) \mid B \subset A\}$ be a set of all the experimental results that can be obtained as protocols (submodels) of observations. Using [28] it is not difficult to prove the following theorem.

Theorem 1. If PR_M is a set of all finite submodels of the subject domain M, then the axiomatic system $S^{\Im}$ is logically equivalent to the set of universal formulas.

We will assume that the condition of the theorem is true for our subject domain and, hence, the axiomatic system $S^{\Im}$ is universally axiomatizable.

8. What is the law?

It is known, that a set of universal formulas $S^{\Im}$ can be reduced to the set of rules (1) by the logically equivalent transformations with literals $A_0, A_1, \dots, A_k$.

$$C = (A_1\&...\&A_k \Rightarrow A_0), k \geq 0, \quad (1)$$

Therefore, we can assume that the axiomatic system $S^{\Im}$ is a set of rules (1).

Thus, the task of discovering a subject domain theory is reduced to discovering rules (1). Let us analyse this task. What can we say about the truth of the axiomatic system $S^{\Im}$ on a set of experimental results PR_M, by using a logical analysis of axioms?

- The rule $C = (A_1\&...\&A_k \Rightarrow A_0)$ is true on PR_M, if its premise is always false on PR_M. We prove in the theorem below that in this case some logically stronger subrule, linking atoms of the premise, is true on PR_M;
- The rule C is true on PR_M if some of its logically stronger subrule that contains only a part of the premise and the same conclusion, is true on PR_M.

Let us clarify logically stronger subrules from which the truth of the rule follows.

Theorem 2 [48][49]. 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 $\{A_{i1},...,A_{ih},A_{i0}\} \subset \{A_1,...,A_k\}$, $0 \leq h < k$, and

$(A_{i1}\&...\&A_{ih} \Rightarrow \neg A_{i0}) \vdash \neg(A_1\&...\&A_k) \vdash (A_1\&...\&A_k \Rightarrow A_0)$;

(2) $(A_{i1}\&...\&A_{ih} \Rightarrow A_0)$, where $\{A_{i1},...,A_{ih}\} \subset \{A_1,...,A_k\}$, $0 \leq h < k$, and

$(A_{i1}\&...\&A_{ih} \Rightarrow A_0) \vdash (A_1\&...\&A_k \Rightarrow A_0)$,

$\vdash$ – provability in a propositional calculus.

Definition 1. A *subrule* of rule C is a logically stronger rule (1) or (2), defined in the theorem 2 for rule C.

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

Corollary 1. If a subrule of rule C is true on PR_M, then the rule C is also true on PR_M.

Definition 2. A *law* on a set of experimental results PR_M is rule C, which is true on PR_M, and none of its subrules is true on PR_M.

Let L be the set of all laws on PR_M. It follows from the logic and methodology of science that hypotheses that are most falsifiable, simplest and contain the smallest number of parameters can be viewed as laws. In our case, all these properties, that are usually difficult to define, follow from the logical strength of the laws. The subrules are:

- logically stronger than the rules and more prone to become false (falsifiable) because they contain weaker premises and, therefore, applicable to larger datasets;
- simpler, because they contain a smaller number of atomic expressions than the rule;
- include a smaller number of "parameters" (the number of atomic expressions may be regarded as parameters "tuning" rules to data).

Why should the laws be most falsifiable, simplest and contain smallest number of parameters? In general, views differ from one author to another. In our case, for the hypotheses of the form (1), we can give a specific answer to this question. Discovery of laws is in fact solving a more relevant task – identifying logically strongest theory, describing our data and providing a probable meachanism of data generation. It can be proved that from a set of laws L the subject domain theory $S^{\Im}$ is inferrable

Theorem 3 [51]. $L \vdash S^{\Im}$.

Therefore, the task of the subject domain theory $S^{\Im}$ discovery is reduced to discovering a set of laws L.

9. Events and their probability

Let us generalize the notion of a law for a probabilistic case. We define a probability on a set of experimental results PR_M and logical expressions. We assume that objects for the experiment are selected randomly from set A. For the sake of simplicity we introduce the discrete probability function on A as a mapping μ: A → [0,1] such that [14] shown in (2).

$$\sum_{a\in A}\mu(a)=1 \text{ and } \mu(a) \neq 0,\ a \in A.$$
$$\mu(B)=\sum_{b\in B}\mu(b),\ B \subseteq A \qquad (2)$$

The discrete probability function μ^n on the product $(A)^n$ will be thereby defined by

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

More general definitions of probability function μ are considered in [14]. Let us define an interpretation of language L on the empirical system M = ⟨A, W⟩ as mapping I: $\Im \to W$. This mapping associates 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, ... } be variables of language L. The valuation ν is defined as a function ν: X → A.

Let us define the probability for sentences of language L. Let U(ℑ) be a set of all atomic formulas of language L of the form P(x_1,…,x_n); ℜ(ℑ) is a set of all the sentences of language L, obtained by the closure of set U(ℑ) relative to logical operations &,v,¬. The formula $\hat{\varphi}$ is defined by νIφ, φ ∈ ℜ(ℑ), where predicate symbols from ℑ are substituted by the predicates from W, that is by interpretation I and variables of the formula φ are substituted by objects from A by the validation ν. The probability η of sentence φ(x_1,…,x_n) ∈ ℜ(ℑ) on M is defined as follows

$$\eta(\varphi) = \mu^n(\{(a_1,\ldots,a_n) \mid M \models \hat{\varphi},\ \nu(x_1) = a_1,\ldots,\ \nu(x_n) = a_n\}),\ \text{where} \models \text{is the truht on M} \qquad (3)$$

10. General notion of law, probabilistic laws on PR_M

Now we revise the concept of law on PR_M in terms of probability. Let us do it in such a way that the concept of the law on PR_M be a particular case of this definition.

The law is true on PR_M rule, whose subrules are false on PR_M. Let us revise the concept of the law on the PR_M. The law is such a true on PR_M rule, which cannot be made simpler or logically stronger without losing the truth. This property of the law "not to be simplified" allows stating the law not only in terms of truth but also in terms of probability.

Theorem 4 [48][49]. For any rule C = (A_1&...&A_k ⇒ A_0), the following two conditions are equivalent:

1. the rule C is a law on PR_M;
2. (a) $\eta(A_0/A_1\&...\&A_k) = 1$ and $\eta(A_1\&...\&A_k) > 0$;
   (b) the conditional probability $\eta(A_0/A_1\&...\&A_k)$ of the rule is greater than the conditional probability of each of its subrules.

This theorem gives an equivalent definition of the law on PR_M in probability terms.

Definition 3. A probabilistic law on PR_M with conditional probability 1 is rule C = (A_1&...&A_k ⇒ A_0) such that:

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

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

We denote a set of all probabilistic laws on PR_M with conditional probability 1 as LP1.

Corollary 2. A probabilistic law on PR_M with conditional probability 1 is a law on PR_M.

Therefore, the task of the subject domain theory $S^{\Im}$ discovery is reduced to a task of discovering all probabilistic laws on PR_M with the conditional probability equal to 1.

The definition of probability (3) is based on the random choice of objects for the experiment. Experiments are "parts" (submodels) of the empirical system $M = \langle A,W\rangle$, and it is not assumed that the truth values of predicates can change during the experiments. As we noted above a more general definition of the probability function μ is given in [14]. This definition includes cases with "noise", when truth-values of predicates can change. We cannot require the complete logical truth of the laws on PR_M for experiments with "noise" and the definition of the law on PR_M should be changed.

Let us consider items 1 and 2 of the theorem 4 from the standpoint of the "not to be simplified" law:

- a law is such a rule, which is true on PR_M, and it cannot be simplified (to be logically stronger) without a loss of the truth.
- a probabilistic law on PR_M with conditional probability 1 cannot be simplified (to be logically stronger) without loosing (decreasing) the value 1 of the conditional probability, so that it became less than 1.

This makes a following general definition of law feasible:

Definition 4. The law is such a rule of the form (1) based on truth, conditional probability, and other estimations, which cannot be made logically stronger without reducing their estimations.

Therefore, we may generalize the definition of the probabilistic law with conditional probability 1 by omitting condition $\eta(A_0/A_1\&...\&A_k) = 1$ from the point (a) of definition 3. The remaining condition (b) expresses the property of the law in the sense of definition 4.

Definition 5. A probabilistic law on PR_M is a rule $C = (A_1\&...\&A_k \Rightarrow A_0)$, $k \geq 0$, such that the conditional probability of the rule $\eta(A_0/A_1\&...\&A_k)$, $\eta(A_1\&...\&A_k) > 0$ is greater than the conditional probability of each of its subrules.

We denote the set of all probabilistic laws on PR_M as LP.

Corollary 3. $L \subset LP$.

Definition 6. A Strongest Probabilistic Law (SLP-rule) on PR_M is a probabilistic law $C = (A_1\&...\&A_k \Rightarrow A_0)$, which is not a subrule of any other probabilistic law.

We define SPL as a set of all SPL-rules.

Proposition 3. $L \subset SLP \subset LP$.

Consider the task of the subject domain theory $S^{\Im}$ discovery in the "noise" conditions.

Definition 7. We say that the "noise" is "saving", if sets of laws LP1 and LP are equal.

The task of subject domain theory $S^{\Im}$ discovery in the presence of noise is, thus, reduced to two tasks (1) evaluation if the noise is "saving"; (2) discovery of set LP. It follows from theorems 3 and 4, corollary 2 and definition 7 that if the noise is "saving", then $LP \vdash S^{\Im}$., and the task of subject domain theory $S^{\Im}$ discovery is reduced to the set LP discovery. In the paper [50] we describe two examples of "saving" "noise" that satisfy the definition 7.

11. The models of predictions and statistical ambiguity problem

The next question about the subject domain theory $S^{\Im}$ discovery is how to use the theory. The main its use is prediction. Now we consider the models of predictions and statistical ambiguity problem for predictions.

One of the major results of the Philosophy of Science is so-called *covering law model* that was introduced by Hempel in the early sixties in his famous article 'Aspects of scientific explanation' [15][16]. The basic idea of this covering law model is that a fact is explained/predicted by subsumption under so-called *covering law,* i.e. the task of an explanation/prediction is to show that a fact can be considered as an instantiation of a law. In the covering law model, two types of explanation/predictions are distinguished: *Deductive-Nomological* (D-N) explanations/predictions and *Inductive-Statistical* (I-S) explanations/predictions. In D-N explanations/predictions, a law is *deterministic*, whereas in I-S explanations a law is *statistical.*

Right from the beginning, it was clear to Hempel that two I-S explanations can yield contradictory conclusions. He called this phenomenon the *statistical ambiguity* of I-S explanations [15][16]. Let us consider the following example of the statistical ambiguity.

As opposed to the classical deduction, in statistical inference it is possible to infer contradictory conclusions from consistent premises.

Suppose that the theory LP makes the following statements:

- (L1): Almost all cases of streptococcus infection clear up quickly after the administration of penicillin;
- (L2): Almost no cases of penicillin resistant streptococcus infection clear up quickly after the administration of penicillin;
- (C1): Jane Jones had streptococcus infection;
- (C2): Jane Jones received treatment with penicillin;
- (C3): Jane Jones had a penicillin resistant streptococcus infection;

It is possible to construct two contradictory arguments from this theory, one explaining why Jane Jones recovered quickly (E), and the other one explaining its negation why Jane Jones did not recover quickly (¬E)

Argument 1		Argument 2	
L1		L2	
C1,C2		C2,C3	
E	[r]	¬E	[r]

The premises of both arguments are consistent with each other. They could all be true. However, their conclusions contradict each other, making these arguments rival ones. So, the set of rules LP may be inconsistent.

Hempel hoped to solve this problem by forcing all statistical laws in an argument to be maximally specific. That is, they should contain all relevant information with respect to the domain in question. In our example, premise C3 of the second argument invalidates the first argument, since law L1 is not maximally specific with respect to all information about Jane in LP. Thus, theory LP can only explain ¬E, but not E. We will return to this example below.

The Deductive-Nomological explanations/predictions of some observed phenomenon G are inferred by the rule:

$$\frac{\frac{L_1,\ldots,L_m}{C_1,\ldots,C_n}}{G}$$

It satisfies the following conditions:

i. $L_1,\ldots,L_m$ are universally quantified sentences (having at least one universally quantified formula), $C_1,\ldots,C_n$ have no quantifiers or variables;

ii. $L_1,\ldots,L_m, C_1,\ldots,C_n \Rightarrow G$;

iii. $L_1,\ldots,L_m, C_1,\ldots,C_n$ is consistent;

iv. $L_1,\ldots,L_m \not\Rightarrow G; C_1,\ldots,C_n \not\Rightarrow G;$

We assume that for deductive-nomological inference of predictions we will use laws from L. Therefore, due to theorem 3, we may infer any prediction that follows from the subject domain theory $S^{\Im}$ [58].

Hempelian inductive-statistical explanations/predictions of some observed phenomenon G are inferred by the analogous rule:

$$\frac{\begin{array}{l}L_1,\ldots,L_m\\ C_1,\ldots,C_n\end{array}}{G}\ [r]$$

where [r] is the probability of inference.

In addition to points i-iv, it satisfies the following Requirement of Maximal Specificity RMS:

v. RMS: All laws $L_1,\ldots,L_m$ are maximal specific.

In Hempel [15][16] the RMS is defined as follows.

An I-S argument of the form:

$$\frac{\begin{array}{l}p(G;F) = r\\ F(\boldsymbol{a})\end{array}}{G(\boldsymbol{a})}\ [r]$$

is an acceptable I-S prediction with respect to a knowledge state K, if the following requirement of maximal specificity is satisfied. For any class H for which the following two sentences are contained in K

$$\begin{array}{l}\forall x(H(x) \Rightarrow F(x)),\\ H(\boldsymbol{a}),\end{array} \tag{4}$$

exists a statistical law p(G;H) = r' in K such that r = r'. The basic idea of RMS is that if F and H both contain the object ***a***, and H is a subset of F, then H provides more specific information about the object ***a*** than F, and therefore law p(G;H) should be preferred over law p(G; F).

For inductive-statistical inference of predictions, we may use probabilistic laws LP. However, in the next sections we present a new definition of a maximum specificity requirement and a coressponding definition of maximum specific rules that solve the problem of statistical ambiguity.

12. Semantic Probabilistic Inference of the Set of Laws L and LP

In this section, we define a semantic probabilistic inference of the sets of laws L, probabilistic laws LP and SLP. This inference also gives us possibility to define maximum specific rules.

Definition 8 [46]**. A** *semantic probabilistic inference* (SP-inference) of some SPL-rule is a sequence of probabilistic laws, which we designate as $C_1 \sqsubset C_2 \sqsubset \ldots \sqsubset C_n$ such that:

$$\begin{array}{l}C_1,C_2,\ldots,C_n \in \text{LP}, C_n - \text{SPL rule}, C_i = (A_1^i \,\&\ldots\&\, A_{ki}^i \Rightarrow G)\,, i = 1,2,\ldots n, n \geq 1\ ,\\ \quad\text{the rule } C_i \text{ is subrule of the rule } C_{i+1},\ \eta(C_{i+1}) > \eta(C_i),\ i = 1,2,\ldots n\text{-}1,\end{array} \tag{5}$$

Proposition 4. Any probabilistic law belongs to some SPI-inference.

Proposition 5. There is a SPI-inference for any SPL-rule.

Corollary 4. For any law from L there is a SPI-inference of that law.

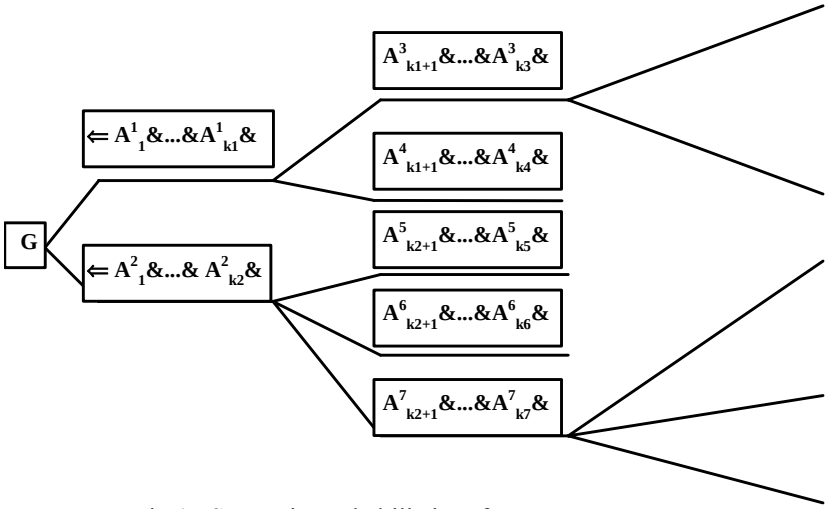

Fig 1. Semantic Probabilistic Inference tree.

Let us consider a set of all SP-inferences of some sentence G. This set constitutes the semantic probabilistic inference tree of sentence G (see fig. 1).

Definition 9. A *maximum specific rule* MS(G) of the sentence G is a SPL-rule of the semantic probabilistic inference tree of sentence G that has maximum value of the conditional probability.

We define a set of all maximum specific rules for any atom G as MSR.

Proposition 6. $T \subset MSR \subset SPL \subset TP$.

13. Requirement of Maximal Specificity and Solution of the statistical ambiguity problem

Now we define the RMS for a probabilistic case. We suppose that class H of objects in (4) is defined by some sentence $H \in \Re(\Im)$ of the language L. Therefore, according to RMS $p(G;H) = p(G;F) = r$ for this sentence. In terms of probability, it means that $\eta(G/H) = \eta(G/F) = r$ for any $H \in \Re(\Im)$ that satisfies (4).

Definition 9. The rule $C = (F \Rightarrow G)$ satisfies the *Probabilistic Requirement of Maximal Specificity* (PRMS) iff:

the equtions $\eta(G/F\&H) = \eta(G/F) = r$ for the rules $C = (F \Rightarrow G)$ and $C' = (F\&H \Rightarrow G)$ follow from: $H \in \Re(\Im)$ and $F(\mathbf{a})\&H(\mathbf{a})$ (in this case the sentence (4) $\forall x(F(x)\&H(x) \Rightarrow F(x))$ is true and $\eta(F\&H) > 0$ due to (2)),

In other words, PRMS means that there is no other sentence $H \in \Re(\Im)$ that increases or decreases the conditional probability $\eta(G/F) = r$ of the rule C by adding it to the premise. See lemma 1 below.

Lemma 1 [52]. If sentence $H \in \Re(\Im)$ decreases the probability $\eta(G/F\&H) < \eta(G/F)$ then the sentence $\neg H$ increases it: $\eta(G/F\&\neg H) > \eta(G/F)$.

Lemma 2 [52]. For any rule $C = (B_1\&...\&B_t \Rightarrow A_0)$, $\eta(B_1\&...\&B_t) > 0$ of the form (2) there is a probabilistic law $C' = (A_1\&...\&A_k \Rightarrow A_0)$ on M which is subrule of the rule C and $\eta(C') \geq \eta(C)$.

Theorem 3 [52]. Any MS(G) rule satisfy PRMS.

Corollary 5 [52]. Any law on M satisfies the PRMS requirement.

Theorem 4 [52]. The I-S inference is consistent for any laws $L_1,\ldots,L_m \in MSR$.

It follows from the theorem, that after discovering a set of all maximum specific rules MSR we can predict without contradictions by using I-S inference.

Let us illustrate this theorem by using the previous example. The maximum specific rules MS(E) and MS(¬E) for the sentences E and ¬E are the rules:

[L1']: 'Almost all cases of streptococcus infection, that are not resistant streptococcus infection, clear up quickly after the administration of penicillin';

[L2] : ‘Almost no cases of penicillin resistant streptococcus infection clear up quickly after the administration of penicillin’.

The rule L1’ have the greater value of conditional probability, then the rule L1 and hence is a MS(E) rule for E. These two rules can’t be fulfilled on the same data.

14. Relational Data Mining and ‘Discovery’ system

This paper reviewed the theory behind the Relational Data Mining (RDM) approach to the knowledge discovery and the ‘Discovery’ system [23, 47] based on this approach. The novel part of the paper is in blending RMD approach, representative measurement theory and current studies on ontology. In this approach, the initial rule/hypotheses generation is task-dependent. More detail about examples of such domain and task specific set of rules/hypotheses are presented in [23] for an initial set of hypotheses for financial time series. For a particular task and a subject domain, the RDM system selects rules that are simplest and consistent with measurement scales (data types). It implements the semantic probabilistic inference and discovers all sets of rules L, LP, SLP, and MSR using data represented as a many-sorted empirical system. In this way a complete and consistent set of rules can be discovered. The system was successfully applied for solving many practical tasks from cancer diagnostic systems, time series forecasting to psychophysics, and bioinformatics (see scientific discovery website [42] for more information).

ACKNOWLEDGEMENTS

The work has been supported by the Russian Federation grants (Scientific Schools grant of the President of the Russian Federation 4413.2006.1 and Siberian Branch of the Russian Academy of Sciences, Integration project No. 1, 115).

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, Muggleton S (1995): Applications of inductive logic programming. Communications of ACM 38 (11):65-70.
[5] Carnap, R., Logical foundations of probability, Chicago, University of Chicago Press, 1962.
[6] 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.
[7] 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).
[8] 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.
[9] 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 Colloquium (1967) 156-172.
[10] Fishbern PC (1970): Utility Theory for Decision Making. NY-London, J.Wiley&Sons.
[11] Flann, N., & Dietterich, T. (1989). A study of explanation based methods for inductive learning. Machine Learning, 4, 187--226.
[12] 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.

[13] 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.
[14] Halpern, J.Y. (1990), `An analysis of first-order logic of probability', Artificial Intelligence 46, pp.311-350.
[15] Hempel, C. G. (1965) Aspects of Scientific Explanation, In: C. G. Hempel, Aspects of Scientific Explanation and other Essays in the Philosophy of Science, The Free Press, New York.
[16] Hempel, C. G.: 1968, 'Maximal Specificity and Lawlikeness in Probabilistic Explanation', Philosophy of Science 35, 116–33.
[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] Kovalerchuk B (1973): Classification invariant to coding of objects. Comp. Syst. 55:90-97, Institute of Mathematics, Novosibirsk. (in Russian).
[21] 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.
[22] Kovalerchuk B, Vityaev E (1998): Discovering Lawlike Regularities in Financial Time Series. *Journal of Computational Intelligence in Finance* 6 (3):12-26.
[23] Kovalerchuk, B., Vityaev, E. (2000), Data Mining in finance: Advances in Relational and Hybrid Methods, Kluwer Academic Publishers, 308 p.
[24] 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.
[25] 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.
[26] 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.
[27] Lebowitz, M. (1986). Integrated learning: Controlling explanation. Cognitive Science, 10.
[28] Mal'tsev A. I. Algebraic systems, Springer, 1971.
[29] Mitchell, T. (1997), Machine Learning, New York: McGraw Hill.
[30] 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.
[31] 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.
[32] Muggleton S. (1999): Scientific Knowledge Discovery Using Inductive Logic Programming, Communications of ACM, vol. 42, N11, pp. 43-46.
[33] Pazzani, M. J. (1990). Creating a memory of causal relationships: An integration of empirical and explanation based learning methods. Hillsdale, NJ: Lawrence Erlbaum Associates.
[34] 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.
[35] Pazzani, M., Kibler, D. (1992). The utility of prior knowledge in inductive learning. Machine Learning, 9, 54-97
[36] Pazzani, M., (1997), Comprehensible Knowledge Discovery: Gaining Insight from Data. First Federal Data Mining Conference and Exposition, pp. 73-82. Washington, DC
[37] Pfanzagl J. (1971). Theory of measurement (in cooperation with V.Baumann, H.Huber) 2nd ed. Physica-Verlag.
[38] Quinlan, J. R. (1989). Learning relations: Comparison of a symbolic and a connectionist approach (Technical Report). Sydney, Australia: University of Sidney.
[39] Quinlan, J. R. (1990). Learning logical definitions from relations. Machine Learning, 5, 239-266.
[40] Samokhvalov, K., (1973). On theory of empirical prediction, (Comp. Syst., #55), 3-35. (In Russian)
[41] 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.
[42] Scientific Discovery website http://www.math.nsc.ru/AP/ScientificDiscovery

[43] 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.
[44] 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).
[45] Vityaev E. (1983). Data Analysis in the languages of empirical systems. Ph.D. Diss, Institute of Mathematics SD RAS, Novosibirsk, p.192. (In Russian)
[46] Vityaev, E.E. (1992), `Semantic approach to knowledge base development: Semantic probabilistic inference', Computer Systems 146, pp.19-49. (in Russian).
[47] Vityaev, E.E., Moskvitin, A.A. (1993), `Introduction to discovery theory: Discovery software system', Computational Systems 148, pp.117-163. (in Russian).
[48] Vityaev E., Logvinenko A. (1995). Axiomatic system testing method, Computational Systems, Theory of computation and languages of specification, (Comp. Syst., #152), Novosibirsk, p.119-139. (in Russian).
[49] Vityaev EE, Logvinenko AD (1998): Laws discovery on empirical systems. Axiom systems of measurement theory testing. Sociology: methodology, methods, mathematical models (Scientific journal of the Russian Academy of Science) 10:97-121. (in Russian);
[50] Vityaev E., Demenkov P. (2003). Empirical Theory Discovery. In: Probabilistic ideas in science and philosophy (Proceedings of the region conference, Novosibirsk, 23-26 sept., 2003), Novosibirsk, pp.86-89.
[51] Vityaev, E., Kovalerchuk, B., Empirical Theories Discovery based on the Measurement Theory. Mind and Machine, v.14, #4, 551-573, 2004.
[52] Vityaev E. The logic of prediction. In: Mathematical Logic in Asia. Proceedings of the 9th Asian Logic Conference (August 16-19, 2005, Novosibirsk, Russia), World Scientific, Singapore, 2006, pp.263-276
[53] Zagoruiko N.G., Elkina V.N. Eds. (1976), Machine Methods for Discovering Regularities. Proceedings of MOZ'76, Novosibirsk. (In Russian)
[54] Zighed, DA (1996): SIPINA-W, http://eric.univ-lyon2.fr/~ricco/sipina.html , Université Lumière, Lyon.