Relational Methodology for Data Mining and Knowledge Discovery

Evgenii Vityaev
Sobolev Institute of Mathematics, SB RAS, Acad. Koptyug prospect 4,
Novosibirsk, 630090, Russia, e-mail vityaev@math.nsc.ru

Boris Kovalerchuk
Computer Science Department, Central Washington University,
Ellensburg, WA, 98926-7520, USA, borisk@cwu.edu

Abstract

This paper analyses capabilities of Machine Learning and KDD&DM methods to perform cognitive processes in the form of discovering the domain theories. The concept of cognition of the domain theory is derived for the reprehensive measurement theory (RMT). We show that a Relational Data Mining approach we proposed previously performs cognition of domain theories in accordance with the RMT and produces the relational methodology for analysis of cognitive capabilities of data mining methods. In this methodology a domain theory includes a metadata ontology. This ontology contains various data types formalized in the first-order logic in accordance with the RMT. To represent the knowledge theory we use the concept of the Logical Empirical Theory that is defined in the paper.

1. Introduction

This paper continues the analysis of Machine Learning and KDD&DM methods capabilities to perform the process of cognition of domain theories started in [1]. A variety of KDD&DM methods have been developed, however their abilities to model the process of cognition of domain theories remain unclear. Can these methods be viewed as a sophisticated way of modelling human cognitive proceses? If the answer is positive at least for some of these methods then it opens wide opportunities for modelling and understanding complex human cognitive processes. To be able to answer theses questions we need to clarify at first what is the cognitive process. The Representative Measurement Theory (RMT) originated by P. Suppes and other scientists at Stanford University [2-4] provides a foundation for answering these questions.

It follows from the RMT that the process of cognition of a domain is stepwise:

1. identify a domain theory as a set of its attributes and laws;
2. determine a set of relations and operations, which constitute the operational sense of the domain attributes and laws;
3. discover all sets of axioms, which are fulfilled for these relations and operations;
4. determine numerical representations of all attributes and laws using the RMT theorems. The theorems determine the numerical representations of the attributes and laws for the corresponding sets of axioms;
5. perform simultaneous scaling procedure for all attributes involved in the discovered laws;
6. derive the structure of all (rescaled) quantities interrelated by the laws for the explored domain theory.

To our knowledge, this description of the process is the most elaborate in modern science. However, this process has several limitations that we discuss later.

The RMT treats the numerical representations of the values and laws as only numerical codes of the algebraic structures representing the operational properties of the values and laws. Thus, the algebraic structures are the *primary representations* of values and laws. The main statements and results of the Measurement Theory are the following [4]:

- numerical representations of quantities and laws of nature are determined by the set of axioms for the corresponding empirical systems, and algebraic systems with certain sets of relations and operations;
- the numerical representations are unique up to certain sets of permissible transformations (such as change in measurement units);
- all physical attributes may be embedded into the structure of physical quantities;
- physical laws are simple because all attributes involved in a law are simultaneously scaled by a special scaling process (there are no KDD&DM methods to discover such laws);

- this axiomatic approach is applicable not only to the physical attributes and laws, but also to many other attributes and laws of other domains (such as psychology).

These are several limitations of the cognitive process outlined above that severely restrict its possible computational implementation:

a) there is no 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 an axiomatic system discovery in the noise conditions;
c) the RMT theorems on the existence of numerical representations are known only for certain axiomatic systems;
d) there is no general method of simultaneous scaling of all the attributes involved in a law.

Let us analyze the cognition process from the KDD&DM viewpoint. Let us divide it into two steps:

Step 1: Identification of the set of relations and operations, the discover of all the axiomatic systems for every attribute and law;

Step 2: Determination of all the numerical representations of all attributes and laws, using the theorems for the Measurement Theory. For any discovered law, a simultaneous scaling procedure should be performed for all the attributes involved in the law. Determination of the structure of all attributes interrelated by laws for the explored domain theory.

Let us define step 1 as performing a Logical Empirical Theory (LET) discovery of the explored domain. Step 2 as performing a Quantitative Empirical Theory (QET) discovery based on the results of the Measurement Theory. According to the methodology of the Measurement Theory the cognition processes should start with a discovery of the Logical Empirical Theory, i.e. from the step 1. The LET represents the qualitative description of the explored domain. A *transformation* of the domain theory from a qualitative to a quantitative state should be performed using the Measurement Theory that is in step 2. Historical developments in the domain theory have been always associated with transformations of this kind.

Let us restrict the domain theory cognition by the step 1, i.e. by the Logical Empirical Theory (LET) cognition. There are reasons for this:

- Step 2 should be performed using the Measurement Theory, not the KDD&DM methods;
- Step 2 follows the historical traditions; it is very convenient for a human to use numbers for attributes and laws. The purely operational and algebraic representation is unacceptable for a human. Nevertheless, we now have tools to operate with the LET, using, for example, the logical programming methods;
- Step 2 places constraints on the cognitive process:
  - there are attributes that have no natural numerical representations. These attributes are structural, partial orders, lattices, graphs, test results, preference relations, etc;
  - there are laws that have no natural numerical representations, such as diagnostic rules, utility functions, trading figures, psychological tests, etc;
- Step 2 yields the *Domain Theory* as a *Quantitative Empirical Theory* represented by the structure of attributes interrelated by the laws. This representation is well elaborated.

We consider the *Logical Empirical Theories* (LET) as a more adequate and modern way of a domain theory representation. The Domain Theory as the LET has no yet a convenient representation, but some representations are currently available, for example, the LET may be represented as an expert system, a logic program or a logical theory.

In compliance with this idea, we have developed the Relational Data Mining (RDM) approach [5,6] to the LET discovery. It integrates the achievements of the Measurement Theory with objectives of Machine Learning and KDD&DM.

The RDM approach is intended to overcome the above restrictions of the step 2. We use the Measurement Theory not as the theory of numerical representations, but conversely, as the theory of correct logical representation of numeric values of quantities and laws. Specifically, according to the Measurement Theory any numerical data type can be transformed into a relational form that preserves relevant properties of a numerical data type. The process of cognition of domain theory in the Relational Data Mining is the following:

i. transformation of all the empirically interpretable information from data into a many-sorted empirical system, i.e., an algebraic system with empirically interpretable sets of relations and operations;
ii. LET discovery for the many-sorted empirical system with presence of noise using a specific RDM method “Discovery” for regularities as logical expressions in the first-order logic with probabilistic estimates [1,5,6].
iii. use of a hierarchy of data types to speed up the search for regularities. The search starts by testing rules based on the properties from the weaker scales and ends up with properties of stronger scales. The number of search computations for weaker scales is much smaller than that for the stronger;

It has been demonstrated that steps (i) and (ii) are applicable to a variety of data types, such as pairwise and multiple comparisons; attribute-based, order, and coherence matrices [5]. These data types can be converted into many-sorted empirical systems. We argue that the current KDD&DM methods utilize only a part of data types information really present in the data. Incorporation

of such information opens new opportunities for increasing the performance of the KDD&DM methods.

For implementing step (iii), we developed a rather general RDM method "Discovery" [5,6]. This method produces, the strongest (in logical sense) LET [1]. Thus, it implements the cognitive process of the Logical Empirical Theory discovery.

In comparison with other KDD&DM methods the Relational Data Mining approach has the following specifics [7]:

I. Any KDD&DM method assumes explicitly or implicitly defined:
   i. data types;
   ii. language to manipulate and interpret data (ontology of particular KDD&DM method);
   iii. class of hypothesis to be tested on data (knowledge theory of particular KDD&DM method);

II. The methodology for Data Mining methods which we want to present consists in considering of different KDD&DM methods from the viewpoint of their Data Types (ontology), Languages and Hypotheses (knowledge theory).

III. In RDM approach we overcome the limitations of particular KDD&DM methods, which are induced by their ontology (data types and languages to manipulate and interpret data) and knowledge theories (hypothesis classes to be tested) by unlimited extension the data type notion and hypothesis classes using the first-order logic, in particular:
   iv. extending the data type notion, using first-order logic and the Measurement Theory for describing various data types;
   v. using any background knowledge for learning and forecasting;
   vi. introducing the notion of Rule Type for defining the hypotheses classes;
   vii. introducing the notion of law-like rules satisfying all the properties of scientific laws: simplicity, maximum refutability and logical generality;
   viii. by developing the DM tool "Discovery", which may discover the knowledge as a set of law-like rules by the specification of the ontology of the method (data types of data and language for manipulate and interpret data), and knowledge theory of the method (rule type of hypotheses to be tested);

2. Representative Measurement Theory and KDD&DM methods invariance

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.

$$\mathbf{A} = <A, P_1,\ldots,P_n, \rho_1,\ldots, \rho_m>$$

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. The relational structure $A = <A, P_1,\ldots,P_n, \rho_1,\ldots, \rho_m>$ is considered along with a numerical of the same type

$$\mathbf{R} = <R,T_1,\ldots,T_n, \sigma_1,\ldots, \sigma_m>$$

where the set R is a subset of Re^m, $m \geq 1$, where 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; every real-value 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 as a numerical system designed as a numerical representation of **A**. The idea of a numerical representation is formalized by the notion of a homomorphism φ: $\mathbf{A} \rightarrow \mathbf{R}$.

A mapping φ: $A \rightarrow Re^m$ is called a (strong) *homomorphism* if:

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

The 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 sequence is the goal: (1) find a numerical representation **R** for a relational structure **A,** (2) prove a theorem that homomorphism φ: $\mathbf{A} \rightarrow \mathbf{R}$ exists; and (3) define a set of all possible transformations f: $\mathbf{R} \rightarrow \mathbf{R}$ of the homomorphism $f\varphi$: $\mathbf{A} \rightarrow \mathbf{R}$ (the uniqueness theorems) for a given the relational structure **A** and numerical representation **R**.

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

Example: A relational structure $\mathbf{A} = <A,P>$ is called a semi-ordering, if for all $a,b,c \in A$ it satisfies the axiom:

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

Theorem: If $\mathbf{A} = <A,P>$ is semi-ordering, then there exists a function U: $A \rightarrow Re$ such that:

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

There are hundreds numerical representations known in the measurement theory. The most popular are several numerical data types. The strongest one is called absolute data type (absolute scale). The weakest numerical data type is nominal data type (nominal scale). In between there

is a spectrum of data types allowing one to compare values with ordering relations, to add, multiply, divide values and so on. The classification of these data types are 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) [4].

Table 1. *Numerical data types.*

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

The transformation group is used for determining the invariance of the laws of nature. 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.

The results of the KDD&DM methods also must not depend on the subjective choice of the measurement units, but it is not the case as usual.

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 [8] (fig. 1):

- W={w}, a training sample;
- X(w) = (x_1,…,x_n), the state of n variables known as attributes for each training example w;
- Y(w), the target function assigning the target value for each training example w;

The KDD&DM method M as a result of the learning on the training set {X(w)}, w∈W generates a rule (model)

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

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

$$J(X(w)) \sim Y(w),$$

where J(X(w)) is a value generated by the rule J, and ~ is the approximate equivalence. The resulting model J can be

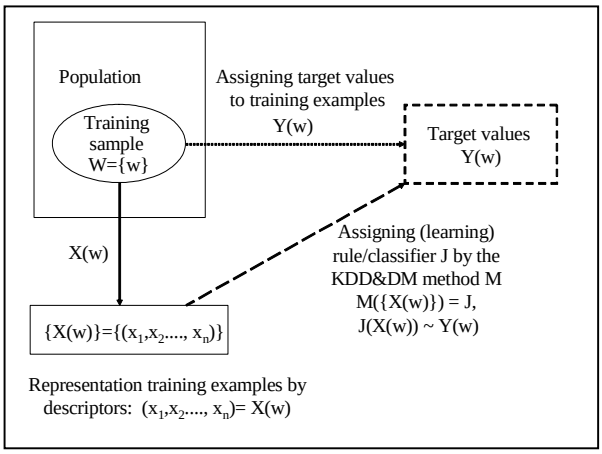

Figure 1. Schematic supervised attribute-based data mining model

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,…,x_n,Y) are determined by the empirical systems A_1,…,A_n,B having the transformation groups g_1,…,g_n,g respectively, then the transformation group G for all the 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∈G

$$J = M(\{X(w)\}) = M(\{g(X(w))\}), \text{ and}$$
$$J(g(X(w))) = g(Y(w)).$$

If the method is not invariant (that is the case for majority of methods) then the resulting rule J or its predictions depends on the subjective choice of the measurement units.

3. Relational methodology

The invariance of a method is closely connected to the interpretability of its results. If the method is non invariant then its results cannot be fully interpretable. In this context interpretation means that concepts used in the method have meaning in the domain, that is they can be explained using concepts known in the domain (domain theory).

The relations and the operations of the empirical systems A_1,…,A_n,B should be interpretable in the system of concepts of the subject domain. If it is not the case, the non interpretable relation or operation is removed from the empirical system and the corresponding scale is reconsidered. For example, an empirical system of the patient temperature measured by a thermometer may include plus operation. The plus operation is interpretable for physical quantities, but may not be interpretable in medicine. Then the plus operation can be removed from

this empirical system and the scale of the corresponding quantity is reconsidered. The order relation for temperature is obviously interpretable in medicine as rising and decreasing of the temperature, and the order relation must be imbedded in the temperature empirical system for medicine.

In contrast a KDD&DM method M is invariant if it uses only information from empirical systems $A_1,\ldots,A_n,B$ as data and produces a rule J, which is logical expression in terms of these empirical systems.

Let us extract an invariant KDD&DM method for some non invariant method M: $\{X(w)\} \rightarrow J$. We will analyze method M from the point of view of its limitations specified in I(i)- I(iii). Let us define an 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 (point I(i)). Let us define the transformation $W \rightarrow A(W)$ of all data into the 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 the method M: $\{X(w)\} \rightarrow J$ let us define the new method ML: $A(W) \rightarrow J$; $ML(A(W)) = M(\{X(w)\}) = J$, where $W \rightarrow A(W) \rightarrow \{X(w)\}$. The method ML uses only interpretable information from data A(W) (see point I(ii)) and produces the rule J by using the method M. Let us analyze the transformation of the interpretable information A(W) in the method ML through the method M into the rule J, and then extract from the rule J some subrule JL, that contains all interpretable information of 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 subrule of the rule J, $A(W) \rightarrow \{X(w)\}$, M: $\{X(w)\} \rightarrow J$. The method MLogic is obviously invariant. If we will consider all possible data for the method M, and all rules JL, that may be produced by the MLogic method, then we will obtain the class of rules {JL} (the class of hypotheses tested, see point I(iii)) of the method M.

As a result of this analysis we obtain: (1) an ontology of the particular KDD&DM method M as an empirical system A(W), and (2) a knowledge theory of that method as the class of rules {JL}. The class of rules {JL} may be compared with the Logical Empirical Theory, discovered by the RDM method “Discovery”. This scheme produces the Relational methodology of analysis of KDD&DM methods.

As a result the metadata ontology in the Relational Methodology will be the first-order logic representation of various data types as it appeared from the Measurement Theory. Thus the knowledge theory may be viewed as a class of rules {JL}.

4. Acknowledgments

This work was partially supported by the Russian Foundation for Basic Research 05-07-90185, Scientific Schools grant at the President of Russian Federation 2112.2003.1, Integration Projects #119 of the Siberian Division of the Russian Academy of Science, NATO (LST.CLG 979815).

5. References

[1] Vityaev, E., Kovalerchuk, B., Empirical Theories Discovery based on the Measurement Theory. Mind and Machine, v.14, #4, 551-573, 2004

[2] Scott, D., Suppes P., (1958), Foundation aspects of theories of measurement, Journal of Symbolic Logic, v.23, pp. 113-128.

[3] 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.

[4] 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.

[5] Kovalerchuk, B., Vityaev, E. (2000), Data Mining in finance: Advances in Relational and Hybrid Methods, Kluwer Academic Publishers, 308 p.

[6] 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.

[7] Scientific Discovery website http://www.math.nsc.ru/AP/ScientificDiscovery

[8] Zighed, DA (1996): SIPINA-W, http://eric.univ-lyon2.fr/~ricco/sipina.html, Université Lumière, Lyon.