Semantic Data Types in Machine Learning from Healthcare Data

Janusz Wojtusiak
Machine Learning and Inference Laboratory
Center for Discovery Science and Health Informatics, George Mason University
Fairfax, VA 22030, USA
jwojtusi@gmu.edu

***Abstract*—Healthcare is particularly rich in semantic information and background knowledge describing data. This paper discusses a number of semantic data types that can be found in healthcare data, presents how the semantics can be extracted from existing sources including the Unified Medical Language System (UMLS), discusses how the semantics can be used in both supervised and unsupervised learning, and presents an example rule learning system that implements several of these types. Results from three example applications in the healthcare domain are used to further exemplify semantic data types.**

Keywords-semantic data types, UMLS, machine learning, healthcare

I. INTRODUCTION

Data are not simply sets of numbers. Each value or symbol has its meaning and it relates to other values or symbols. In healthcare, data can represent patient conditions, treatments, socioeconomic status, genetic information, financial and management decisions, and so on. Every value represents a concept that is related to other concepts represented within that data. Many of these relationships are known by medical and healthcare experts and often coded in hierarchies, ontologies and terminologies; thus, they are computationally processible by machine learning methods.

Despite such richness in domain knowledge, very few machine learning methods are capable of handling semantic information, therefore the problem stays outside of the main focus of the ML community. Most researchers contribute to the theory and practice of using ML systems, which handle all data as numbers and are designed solely to achieve high predictive accuracy disregarding the meaning of the data. Among the best widely used supervised methods in this class are Support Vector Machines (SVM), logistic regression, and random forests.

On the other hand, semantic information is frequently used in Natural Language Processing (NLP), including ontology-based methods for analyzing clinical notes, linking clinical concepts, and using medical terminologies.

This paper argues for the use of semantic information that describes data. It describes semantic data types that can be used by machine learning methods and presents their examples from the healthcare domain. The Unified Medical Language System (UMLS) is the presented example of a medical ontology from which semantic information can be derived. The paper also presents the AQ21 system, which implements several of the discussed semantic data types. Finally, a set of example healthcare application problems is presented to illustrate the use of semantic data types.

The presented data types and reasoning methods are exemplified by their roles in AQ learning, however the presented concepts are general and applicable to any methods that aim at handling complex data.

II. THE NEED FOR SEMANTIC INFORMATION

Although Stevens' attribute types are widely recognized in machine learning and statistics, most machine learning programs do not have mechanisms for appropriately reasoning with all of them, not to mention reasoning with the larger set of types and more advanced structures described in this paper. This fact can be attributed mainly to two reasons: (1) most state-of-the-art ML algorithms are based on numerical methods (i.e., inversion of matrices, distance measures between vectors, etc.) thus are inappropriate for handling semantic information; and (2) the algorithms are designed to efficiently analyze very large amounts of data to achieve high predictive accuracy, and do not concern situations when little data are available nor when criteria other than accuracy is important. While this paper focuses mainly of supervised learning methods the above are true also for unsupervised learning.

On the other hand, some authors consider several other criteria for evaluating ML methods. A summary of some of the criteria has been presented by Wojtusiak and Baranova [19] and includes: accuracy, transparency, acceptability, efficiency, and exportability. Using semantic information that describes data can help on all the criteria. Additional background knowledge may help guide the generalization process, thus learning may achieve better accuracy of created models, and at the same time better performance (i.e., due to smaller search space to be explored constrained by semantics). In healthcare applications, acceptability of ML methods, and models created by ML is very low. The use of semantic information may greatly improve the acceptability (a form of face-validity of model) by making it consistent with common expert knowledge. Very often, models are rejected despite their good accuracy because they contradict

what is "obvious" to domain experts or confirm the obvious without providing anything new.

While the use of different data types by machine learning software is not new, most methods only implement a limited set of types that are targeted at specific application or form of learning. Many machine learning software, for example, C4.5 [12], RIPPER [2], and CN2 [1] recognize only nominal (discrete) and continuous (for all numeric types) attributes. The widely used program Weka, which includes several learning methods [15], additionally supports the type "date." To the author's best knowledge, the only integrated multi-operator systems that support a wider set of attribute types are INLEN and its successor VINLEN.

Structured attributes were first implemented in the INDUCE relational learning program, (e.g. [6]). Kaufman and Michalski presented a detailed study of their use in AQ learning [5]. In the literature, structured attributes are now commonly used and often referred to as hierarchical attributes. They are used in several machine learning and data mining systems, for example in SAMUEL [3] and in the commercial database Oracle. The importance of this attribute type was recognized by a workshop on dealing with structured data in machine learning and statistics at the European Conference on Machine Learning, ECML-2000.

An approach to incorporate structures in output attribute when learning multiclass problem is discussed by several authors, including Sahbi and Geman who proposed a hierarchical model that combines multiple support vector machines. However, according to the literature, there is currently no SVM-based method that allows for directly using hierarchical input attributes without special pre-processing of data.

The following section describes a methodology for reasoning with different data types specifically targeted for medical/healthcare applications. The list of data types and semantic relationships are, however not exhaustive and can be extended when needed by a specific application.

III. Semantic Data Types

Semantic data description includes information about attribute types, inter-attribute relationships, value aggregation semantics, data transformations, and meta-values. The basis for the discussion below is Attributional Calculus, a logic and representation system for machine learning proposed by Michalski [8]. The calculus acknowledges only selected attribute types and does not discuss other important issues such as inter-attribute relationships. The following sections describe types of semantic data information that can be used in machine learning. Additionally, Table 1 summarizes the use of semantic information.

A. Attribute Types

Attributional Calculus recognized nine main attribute types with their semantic and syntactic interpretation [9]. These basic attribute types are: nominal, ordinal, cyclic, structured, set-valued, interval, ratio, absolute, and compound. This paper introduces additional attribute types graph and set-defined with its subtypes, which are specifically important when reasoning with medical data. This section outlines all of the above attribute types, starting with those having discrete domains and followed with attributes with numeric domains.

Nominal (a.k.a. categorical), is an attribute type whose domain is an unordered set. Examples of nominal attributes include, Blood Type with domain {0+, A+, B+, AB+, 0-, A-, B-, AB-}, and Gender with domain {male, female, other}.

Ordinal, is an attribute type with a domain that is a totally ordered finite set. Examples of ordinal attributes include, Student's Grades with domain {F, D, … B, B+, A-, A, A+ }, and Difficulty Level with domain {low, medium, high}.

Cyclic, is an attribute type whose domain is a finite cyclically ordered set. Examples of cyclic attributes include, Hours of the day, Days of the week, Months of the year, Signs of the zodiac, and Time zones.

Structured (a.k.a. hierarchical), is an attribute type whose domain is a hierarchically ordered set. Examples of structured attributes include Diagnosis codes (i.e. ICD-9 codes – International Classification of Diseases, version 9) whose values form a hierarchy, Shapes with domain {polygon, rectangle, square, hexagon, oval, circle, and ellipse}, Body parts, and Drugs. Hierarchies usually follow Is-A relationship (i.e., colon cancer is-a cancer), and are very commonly used in healthcare, with the majority of terminologies including some forms of hierarchies. An illustration of generalization with a structured attribute is presented at the end of this section.

Graph, is an attribute type whose domain is a set whose values are organized into a directed graph G=(V, E), or G=(V, E, Ev) where V is a finite set of nodes, E is a set of edges between nodes, and Ev is an optional set of edge values. Each edge includes information about relationship type between nodes. An example is a floor plan of a hospital building. The set V includes all rooms in the building, the edges E define if there is a passage between the rooms, and the optional set Ev defines the distance (or walking time) when following an edge. Another example is an attribute whose values describe body parts and edges define their proximity to each other.

Set, is an attribute type whose domain is the power set of a base set P(S), and its values are subsets of the base set. Examples of a set attribute include Diagnoses of a patient (a number of comorbidities may vary among patients, but all diagnoses come from a pre-defined set), and Items one bought in a department store (a widely-known *itemset* in data mining). It is possible to represent a set attribute with several binary attributes (yes/no for each possible value), however it may be impractical when the number of potential values is very large.

For example, there are over 14,000 different ICD-9 codes; thus, several thousand attributes would need to be defined in the data to represent them. It is also not feasible to code diagnoses as $Diagnose_1$, $Diagnose_2$, … $Diagnose_k$ because the number of distinct diagnoses for patients may vary up to over a hundred within one 3-month period (i.e., for nursing home patients). Thus, representation of diagnoses for patients with multiple comorbidities is an elegant and efficient solution.

Set-defined, is an extension of the set attribute type. Set elements are provided with additional structure corresponding to attribute types discussed above (nominal, ordinal, cyclic, structured, and graph). The following are specific types of set-defined attribute types. Specifically, *Set-nominal* is the same as *set* attribute type; Set-*ordinal* consists of values that are linearly ordered; *Set-structured* consists of values that form a hierarchy; and *Set-graph* consists of values that are nodes in a graph.

Interval, is an attribute type whose domain is a set of real numbers with a defined difference. An example of an interval attribute is Temperature in Fahrenheit degrees. It is meaningful to say the patient's temperature increased by 5 degrees, but it is not meaningful to say patient's temperature is twice as high as it should be (because zero is not well-defined for Fahrenheit degrees).

Ratio, is an attribute type whose domain is a set of real numbers with a defined difference and product. For example, Height is a ratio attribute because the statement patient A is twice as tall as patient B is meaningful. It is also possible to distinguish *Ratio-integer* attribute type (*integer* or *count* for short) whose domain is the set of natural numbers. For example, WBC count (White blood cells count) is an integer attribute.

Cyclic-ratio, is an attribute type whose domain is an infinite cyclically ordered set. An example of cyclic-continuous attribute is Angle with domain $\{x \in R: 0 \leq x < 360\}$.

Absolute, is an attribute type whose domain is a set of integers, but neither addition nor multiplication are meaningful. An example is a Social Security Number. Although we can compare two numbers and decide which one is higher, saying that one SSN is higher than another SSN by 10,000 does not have any meaning.

Compound, is an attribute whose domain is the Cartesian product of the domains of its constituent attributes. A compound attribute is used to characterize an object or parts of an object in terms of constituent attributes that apply only to this object or to individual parts. Such a characterization is done by listing values of constituent attributes for this object or for a specific part, but without stating attribute names. Compound attributes can be described in analogy to Classes in object-oriented programming. For example, "Weather" can be used as a compound attribute whose values are lists of properties typically used to characterize weather. For example, attributional calculus allows one to create an expression "Weather = sunny & humid", where "Weather" is the compound attribute and "sunny" and "humid" are values of its constituent attributes. As one can see, this expression directly corresponds to an equivalent natural language statement "The weather is sunny and humid", and appears to be more natural to people than a standard logic expression such as "The weather-type is sunny and the humidity is yes." Compound attributes thus facilitate natural induction.

Example: Coding Diagnoses

To demonstrate the difference between some of the attribute types the following example shows coding of patient's diagnoses.

Starting with the simplest data type, nominal, we can define attributes corresponding to *Primary diagnosis* (with a domain being a set of potential diagnoses), *Secondary diagnosis* (with a domain being a set of potential diagnoses), *Tertiary diagnosis* (with a domain being a set of potential diagnoses), and so on. The difficulty is that the total set of diagnoses are unknown and may vary greatly among patients, thus a potentially large number of missing values will need to be included in the data. Also, the order of diagnoses is arbitrary.

The same method of designing attributes can be used when including additional information about the hierarchy of diseases. The use of structured attributes may potentially lead to better generalization but both aforementioned problems remain.

In order to avoid problems with an unknown number of diagnoses per patient, and the very large total number of potential diagnoses, *set* attributes can be used. For a given patient an attribute *Diagnoses* may take a value {type II diabetes, colon cancer, pneumonia} indicating that the patient has been diagnosed with all three conditions within a specific time period.

Finally, diagnoses can be encoded as a *Set-structured* attribute with values corresponding to specific diagnoses (i.e. as ICD-9 codes), or *Set-graph* attribute with values corresponding to diagnoses in a medical ontology (i.e. concepts in SNOMED-CT).

B. Aggregated vs. Individual Values

Data processed by the majority of ML software are in the form of attribute-value pairs, and organized in data tables. An individual example (i.e., corresponding to one patient) is given as a set of values:

$$(v_1, v_2, \ldots v_k)$$

In healthcare applications, the individual patient data are often not available and cannot be shared due to privacy and other issues [19]. Instead, only aggregated summaries describing groups of patients are available. Such aggregated summaries are also one of the most common forms in which results are presented in medical publications [20].

Aggregated data are the most common way of describing cohorts of patients [17][19][20]. Aggregated

values of attributes are given in the form of pairs (μ_A, σ_A), where A is a measured attribute, and μ_A and σ_A denote its mean and standard deviation measured over a group for which the aggregation was done. Given that, means and standard deviations for several parameters are available and each group can be described by an *aggregated example* given as:

$$((\mu_{A_1}, \sigma_{A_1}), (\mu_{A_2}, \sigma_{A_2}), \ldots, (\mu_{A_k}, \sigma_{A_k}))$$

For non-numerical attributes, a typically used aggregated form lists the frequencies of values in a group explicitly showing the distribution of examples. For example, a group of patients may include 70% White, 17% Black, 2% Asian patients, and so on.

The learning problem from aggregated data is to create models for categorizing individual patients when no individual training data are available. In other words, the data are given as a set of aggregated examples, and the goal is to classify individual examples.

Semantic attribute types outlined in the previous section extend to aggregated data. After aggregation, a *nominal* attribute becomes *aggregated-nominal* attribute, a *structured* attribute becomes *aggregated-structured* attribute, a *ratio* attribute becomes *aggregated-ratio* attribute, and so on. This is because semantically the relationships between values on an individual level are the same; only instead of individual values their distributions over a cohort of patients are available.

C. Relationships Between Attributes

The previous sections discussed types of relationships that bind together values of a single attribute. However, an equally important class of relationships span between attributes.

Attribute-attribute relationship, binds together two attributes, meaning that for a given example in the data a value of one attribute is related to a value of another attribute in that example. Examples of attribute-attribute relationships include: diagnosis-treatment pair, in which it is known that a treatment in one of the data columns corresponds to a diagnosis on other specific column. This type of relationship is partially equivalent to causal relationships modeled by Bayesian networks and somewhat corresponds to links between nodes.

Value-value relationship, is defined between specific values of different attributes and may not be present in a specific example (only when appropriate values are present). This type of relationship is particularly important when it spans between values of two structured attributes and can be used to constrain the generalization process [11].

D. Data transformation (constructive induction)

Semantic data types allow for specifying allowed operations between attributes by changing the problem representation space, also known as search space, which is the set of all possible problem solutions. Designing suitable representation space for a given optimization or learning problem is one of the most important and challenging tasks.

The original representation space provided to a machine learning or data mining system may be inadequate for performing the desired task for concept learning, pattern discovery, optimization, etc. Constructive Induction (CI) methods automatically create new representation spaces based on the original representations. The new representations allow the determination of relationships that cannot be represented in the original spaces. New representations are created by removing attributes irrelevant to the considered problem through modifying domains of attributes (for example by discretizing numeric attributes), and by creating new attributes.

Attribute semantic types define allowable operations and transformations for constructing new attributes. For example, for numeric attributes only two operations are allowed: count (which assigns number of attributes that have a specific value in an example), and equality (which compares values of two attributes).

E. Meta-values

In addition to regular values specified in the attribute domain there is a need to semantically encode missing values. The semantic meaning is carried out by three *meta-values* [10]. These values represent possible responses to questions requesting the attribute value for a given entity, specifically: “don’t know” (denoted by “?”), “not applicable” (denoted by “NA”), and “irrelevant” (denoted by “*”). While the value “?” represents lack of knowledge, values “NA” and “*” constitute domain knowledge that is communicated to the system by an expert.

The meta-value “don’t know” is given to an attribute whose value for a given entity exists but is unknown for whatever reason; for example, it has not been measured or has not been recorded in the database.

The “not-applicable” value is given to an attribute that is not applicable to a given entity. For example, the attribute Pregnant {Y, N} is not applicable to male patients but it is applicable to female patients.

Finally, the “*” value is assigned to an attribute that can be considered irrelevant to the problem at hand. For example, the attribute “shoe size” can be viewed irrelevant to the problem of learning people’s education level from their other attributes.

Table 1: Summary of semantic data types.

Attribute Type	Domain	Structure of values	Transformations	Inter-attribute relationships	Typical Aggregation	Output structure	Example
Nominal	Symbolic	Unordered set	Count, equality	A-A, V-V	Proportion	Independent models	Gender
Ordinal	Symbolic	Ordered set	Count, equality	A-A, V-V	Proportion	Sequential classifier	Grade
Cyclic	Symbolic	Cyclically ordered	Count, equality	A-A, V-V	Proportion	Independent models	Day of the week
Structured	Symbolic	Hierarchy	Count, equality	A-A, V-V	Proportion	Hierarchical classifier	Shapes
Graph	Symbolic	Directed graph	Count, equality	A-A, V-V	Proportion	Independent models	Roadmap
Set	Symbolic	Set of values	Count, equality, inclusion	A-A, V-V	Proportion	Set of independent models	Treatments
Set-ordinal	Symbolic	Set of ordinal values	Count, equality, inclusion	A-A, V-V	Proportion	Set of sequential models	Recurring events
Set-structured	Symbolic	Set of structured values	Count, equality, inclusion	A-A, V-V	Proportion	Set of hierarchical models	Diagnoses
Set-graph	Symbolic	Set of graph nodes	Count, equality, inclusion	A-A, V-V	Proportion	Set of independent models	Locations on a map
Interval	Symbolic	Linearly ordered	y=x+b, equality, greater	A-A	Mean +/- standard deviation	Regression model	Temperature
Ratio	Numeric	Linearly ordered	y=ax+b, equality, greater	A-A	Mean +/- standard deviation	Regression model	Height
Cyclic-ratio	Numeric	Cyclically ordered	y=ax+b, equality, greater	A-A	Mean +/- standard deviation	Cyclic regression model	Angle
Absolute	Numeric	Linearly ordered	Equality, greater	A-A, V-V	N/A	Independent models	SSN
Compound	Numeric	Defined by constituent attributes	Defined by constituent attributes	Defined by constituent attributes	Defined by constituent attributes	Defined by constituent attributes	Weather

IV. Unified Medical Language System

The Unified Medical Language System (UMLS), maintained by the US National Library of Medicine, contains a large collection of medical concepts, terms, and relationships from more than 160 source terminologies/ontologies such as SNOMED-CT, ICD (multiple versions), RxNorm, and LOINC [7]. It has three major components: Metathesaurus, Semantic Network, and SPECIALIST Lexicon. The 2012AA version of the UMLS contains more than 2.7 million concepts (CUIs) and 10.8 million unique concept names (AUIs) from source terminologies. The UMLS establishes connections between biomedical terminologies/ontologies by assigning a single concept identifier (CUI) to concept names from various terminologies that have the same meaning. The mappings among these vocabularies allow computer systems to translate data among the various terminology systems. The rich relationships among concepts in the UMLS also provide a solid foundation for reasoning the medical knowledge.

The extreme size of UMLS requires selective methods for loading parts of the ontology. One method, presented by Min and Wojtusiak [11] follows relations within a neighborhood of concepts present in the data. The method derives parent concepts that are not more distant that a value specified by a parameter, and have at least two children in the analyzed data. The relationships are used to derive parent concepts and relationships for both internal attribute semantic types, and inter-attribute relationships.

V. AQ21

Many algorithms are available for inducing rules from data. AQ21 [18] is a rule learning system whose goal is to implement methods for reasoning with complex healthcare data, including reasoning with semantic data types described in this paper. To date, the system implements nominal, ordinal structured, graph, interval, ratio, absolute, and

compound attribute types and semantic network for representing inter-attribute (value-value) relationships. Other data types are currently being implemented with the main focus on set and set-structured types needed to represent patients with multiple co-morbidities.

AQ learning works in two main stages: rule construction and rule optimization. At the core of the first stage is a star generation algorithm which creates multiple generalizations (in the form of attributional rules), called *stars*, of a selected positive example that do not cover negative examples. A combination of rules selected from one or more stars is used as a generated hypothesis (set of rules). Rule optimization uses a user-defined set of statistical and transparency-based criteria to fine-tune learned rules. At both stages of learning AQ21 uses semantic information describing data.

VI. Example Application Areas

The AQ21 system has been successfully applied to several healthcare problems including prediction of claims payments [21], comparative effectiveness of treatments and medications [16], and differential diagnosis [17][19].

The first two application areas are based on individual data supplemented with semantic information in the form of the attribute types. In the claims prediction problem, AQ21 outperformed compared methods [4], which is partially attributed to the use of semantic data types and imbalanced nature of the datasets.

The differential diagnosis application area was based on aggregated data derived from published manuscripts. The method has been applied to diagnoses of polycystic ovary syndrome, and metabolic syndrome.

VII. Conclusions

This paper discussed semantic data types that define intra- and inter-attribute structures and relationships, as well as aggregated data and meta-values. The use of semantic data types is particularly important in healthcare, which is very rich in semantic information available to potential users.

The majority of the methodology has been implemented in the AQ21 system, however additional work is needed to validate and further extend the methodology.

Because the focus of this work is on describing semantic data types, a detailed experimental evaluation is out of scope. Such a comparison will require datasets with available semantic information that is typically not present in those used in machine learning benchmarking.

Acknowledgment

The author thanks Chris Jose for her edits. Current activities of the Machine Learning and Inference Laboratory are supported in part by the National Institute for Standards and Technology, the Department of Veterans Affairs, the Mason-Inova fund, and Robert Wood Johnson Foundation.

References

[1] Clark, P. and Niblett, T., "The CN2 Induction Algorithm," *Machine Learning* 3: pp. 261-289. 1989.

[2] Cohen, W., "Fast Effective Rule Induction," *Proceedings of the 12th International Conference on Machine Learning*. 1995.

[3] Grefenstette, J.J., "The Evolution of Strategies for Multiagent Environments," *Adaptive Behavior*, 1, 1, 65-90, 1992.

[4] Irvin, K., Ngufor C., Wojtusiak, J., "Comparison of Classification Learning Methods for Medical Claims Payments," *American Medical Informatics Annual Symposium*, 2012.

[5] Kaufman, K. and Michalski, R.S., "A Method for Reasoning with Structured and Continuous Attributes in the INLEN-2 Multistrategy Knowledge Discovery System," *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (KDD-96), Portland, OR, pp. 232-237, August, 1996.

[6] Larson, J. and Michalski, R.S., "Inductive Inference of VL Decision Rules," *Invited paper for the Workshop in Pattern-Directed Inference Systems*, Hawaii, and published in SIGART Newsletter, ACM, No. 63, pp. 38-44, June 1977, May 23-27, 1977.

[7] Lindberg DAB, Humphreys BL, McCray AT. *The Unified Medical Language System*. Meth Inform Med. 1993;32:281-91.

[8] Michalski, R. S., "ATTRIBUTIONAL CALCULUS: A Logic and Representation Language for Natural Induction," *Reports of the Machine Learning and Inference Laboratory*, MLI 04-2, George Mason University, Fairfax, VA, April, 2004.

[9] Michalski, R. S. and Wojtusiak, J., "Semantic and Syntactic Attribute Types in AQ Learning," *Reports of the Machine Learning and Inference Laboratory*, MLI 07-1, George Mason University, Fairfax, VA, 2007.

[10] Michalski, R. S. and Wojtusiak, J., "Reasoning with Missing, Not-applicable and Irrelevant Meta-values in Concept Learning and Pattern Discovery," *Journal of Intelligent Information Systems*, 39, 1, 141-166, Springer, 2012.

[11] Min, H., Wojtusiak, J., "Clinical Data Analysis using Ontology-guided Rule Learning," Seconnd International Workshop on Managing Interoperability and compleXity in Health Systems, MIX-HS, 2012, (submitted).

[12] Quinlan, J.R. *C4.5 Systems for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

[13] Sahbi, H., Genam, D. "A Hierarchy of Support Vector Machines for Pattern Detection," *Journal of Machine Learning Research*, 7, 2087-2123, 2006.

[14] Stevens, S.S., "On the Theory of Scales of Measurement," *Science*, 103, pp. 677-680, 1946.

[15] Witten, I.H. and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Third Edition, Morgan Kaufmann, 2011.

[16] Wojtusiak, J. and Alemi, F., "Analyzing Decisions Using Datasets with Multiple Attributes: A Machine Learning Approach," *Handbook of Healthcare Delivery Systems*, CRC Press, 2010

[17] Wojtusiak, J., Michalski, R. S., Simanivanh, T. and Baranova, A. V., "Towards application of rule learning to the meta-analysis of clinical data: An example of the metabolic syndrome," *International Journal of Medical Informatics*, 4, 1, pp. 43-54, 2009.

[18] Wojtusiak, J., Michalski, R. S., Kaufman, K. and Pietrzykowski, J., "The AQ21 Natural Induction Program for Pattern Discovery: Initial Version and its Novel Features," *Proceedings of The 18th IEEE International Conference on Tools with Artificial Intelligence*, Washington D.C., November 13-15, 2006.

[19] Wojtusiak, J. and Baranova, A. V., "Model Learning from Published Aggregated Data," *Learning Structure and Schemas from Documents*, Studies in Computational Intelligence, 375, 369-384, 2011.

[20] Wojtusiak. J., Irvin, K., Birerdinc, A., Baranova, A., "Using Published Medical Results and Non-homogenous Data in Rule Learning," *Proceedings of the International Conference on Machine Learning and Applications*, Honolulu, HI, December 2011.

[21] Wojtusiak, J., Ngufor, C., Shiver, J., Ewald, R., "Rule-based Prediction of Medical Claims' Payments: A Method and Initial Application to Medicaid Data" *Proceedings of the International Conference on Machine Learning and Applications*, Honolulu, HI, December 2011.