The Natural Induction System AQ21 and Its Application to Data Describing Patients with Metabolic Syndrome: Initial Results

Janusz Wojtusiak[1], Ryszard S. Michalski[1], Thipkesone Simanivanh[2], and Anna V. Baranova[2]
(1) Machine Learning and Inference Laboratory, (2) Molecular and Microbiology Department
George Mason University
jwojt@mli.gmu.edu, tsimaniv@gmail.com, abaranov@gmu.edu

Abstract

This paper briefly describes the AQ21 learning system that implements a simple form of natural induction, an approach to learning that generates hypotheses in forms resembling natural language descriptions, and by that easy to understand and interpret. The system was applied to the analysis of aggregated data obtained from non-invasive tests performed on different groups of patients with metabolic syndrome. The discovered patterns were very simple and were evaluated by an expert as potentially medically significant.

1. Introduction

In many areas of application, in particular, in medicine, computer-generated knowledge must be not only accurate but also understandable and interpretable by people. Most machine learning research, however, has given primary attention to predictive accuracy of the learned knowledge, and lesser attention to its understandability.

The goal of achieving both high accuracy and knowledge interpretability has led us to the development of the *natural induction* approach to machine learning [7] that seeks inductive hypotheses in forms natural to people, such as natural language-like descriptions.

This paper provides a brief description of an early version of the natural induction system AQ21, implemented in our laboratory, and its experimental application to an important medical problem. The problem is to determine patterns in patients with *metabolic syndrome* (MS) from aggregated data about groups of such patients. Metabolic syndrome represents a cluster of significant health risk factors that are on the rise in many countries. They include obesity, insulin-resistance, hypertension, elevated triglycerides, as well as decreased level of high-density lipoprotein cholesterol ("good cholesterol"). The presence of any three of these factors increases the risk of cardiovascular disease and predisposes a patient to the development of type 2 diabetes and its complications. Recently, *nonalcoholic fatty liver disease* (NAFLD) and its more aggressive form, *nonalcoholic steatohepatitis* (NASH), have come to be regarded as the hepatic manifestation of MS [4]. Additionally, MS plays a well-recognized role in the development of the obstructive sleep apnea, erectile dysfunction, polycystic ovary syndrome, and malignant tumors [20]. The problem is significant because about 47 million residents of the United States have metabolic syndrome [15]. Thus, discovering valid, simple, and useful clinical patterns characterizing MS is a significant diagnostically goal.

In this study, we used aggregated data from clinical journals because obtaining individualized data about specific patients turned out to be difficult as they are protected by strict privacy laws. Because the AQ21 learning program (as most of machine learning programs) was developed to learn from concept examples representing single entities (in this case patients), not from aggregated data; our first task was to develop a methodology for reasoning with aggregated data. The method presented here represents our initial solution to this problem.

2. AQ21 as Natural Induction Laboratory

AQ21 deeply integrates a wide range of features that are either not present in other programs or present only individually. Because it can perform many different functions and generate different kinds of descriptions and patterns, AQ21 can be viewed as a laboratory for natural induction, rather than as a single program (e.g. [18], [19]). A brief overview of its features, especially those relevant to our study on metabolic syndrome, is presented here.

AQ21 seeks descriptions in different forms that generalize examples of given concepts and optimize a multi-criterion measure of description quality (defined by the user). Descriptions are expressed as attributional rules, whose form used in this study is (1):

CONSEQUENT <= PREMISE |_EXCEPTION
: ANNOTATION (1)

where *CONSEQUENT* and *PREMISE* are *attributional statements* that are conjunctions of *attributional conditions*, and *EXCEPTION* (optional) stands for either an e*xception clause* (which is also an attributional statement) or a list of examples constituting exceptions to PREMISE. *ANNOTATION* lists statistical and other information about the rule, such as numbers of positive and negative examples covered, the rule's quality, its complexity, etc. Here is an example of such a rule:

[PlanToDo= run_experiments]
<= [Day= weekend: 38,131] &
[Weather = rainy and cold: 22, 5] &
[Available_workstation_clock_speed >= 2GB/s: 50,5] &
[Available_lab = lab1 or lab3: 43,57]
|_ [Server_malfunction: 0,3] : p=27, n=1, Q=0.9

which can be paraphrased:

> *The plan to do is to run_experiments, if it is weekend, the weather is rainy and cold, the available workstation has the clock_speed at least 2GB/s, and the available laboratory is lab1 or lab2, except for when there is a server malfunction.*

The value "weekend" is a higher-level value of a structured (hierarchical) attribute "Day", the attribute "Weather" is a *compound attribut*e that takes a conjunction of values. The pairs of numbers within conditions represent numbers of positive and negative examples satisfying these conditions. The entire rule is satisfied by 27 positive and 1 negative examples, as indicated by "p=27, n=1" in the ANNOTATION part, and its quality, defined by (2), is 0.9. Notice that the rule closely corresponds to its equivalent natural language interpretation.

Attributional rules learned by AQ21 may include other types of linguistic constructs than those exemplified above, such as *count attributes* that express the number of statements being true or the number of properties satisfying a given condition in the training data (they generalize quantifiers in standard logic), *arithmetic expressions*, *provided-that clauses*, and other forms resembling natural language descriptions.

As shown above, an attributional rule employs a richer representation language than a typical *elementary rule* learned by most rule learning programs, in which conditions are limited to the form (attribute relation attribute-value).

Before we describe further details of AQ21, let us briefly relate it to other rule learning methods. Among recent methods are those that use rough-set theory approach [11], and those applying evolutionary computation [14]. Some recent papers also describe methods partially based on the AQ approach, e.g., [17]. All these methods, as well as earlier ones, such as RIPPER [3], CN2 [2] and C4.5 [13], are significantly different from AQ21, both in the types of rules they can learn (only elementary rules) and in the way they learn them, typically in the top down-fashion, rather than bidirectionally, as in AQ21.

2.1. Top Level Learning Algorithm

AQ21 integrates three basic learning modes. One mode, called TF ("Theory Formation"), learns complete and consistent hypotheses with regard to the training data. The second mode, called ATF ("Approximate Theory Formation"), learns approximated hypotheses that may be partially inconsistent and/or incomplete. The third mode, called PD ("Pattern Discovery"), seeks optimized patterns that capture strong regularities in data. Figure 1 presents the top level of algorithm for all three modes.

```
RS = null
While P is not empty
  Select a seed example, p, from P
    Generate a star or an approximate star G(p, N)
    Select the best k rules from G according to LEF,
       and include them in RS
 Remove from P all examples covered by the selected
    rules
Optimize rules in RS
Assemble a final hypothesis, a set of alternative
   hypotheses, or patterns from all rules in RS
```

Figure 1: Top-level learning algorithm in AQ21.

The input to the algorithm is a set of positive examples, P, and a set of negative examples, N, of the concept to be learned, and a multicriterion quality measure, LEF (Lexicographic Evaluation Functional), of the generated hypotheses. In the case of learning multiple concepts, negative examples are examples of other concepts than the one being learned. The key part of the algorithm is the generation of a *star* G(p, N), for the given seed, p, against the set of negative examples, N. In TF mode, a star is a set of maximally general consistent attributional rules that cover the seed but do not cover any negative examples, and ATF and PD modes, the rules may be partially inconsistent [6].

The above algorithm applies to all three modes, but in ATF and PD modes at each step of star generation it optimizes rules according to a rule *quality measure*, Q(*w)*, instead of generating consistent rules. Q(*w*) defines the desirable tradeoff between rule completeness, denoted *compl*, and confidence gain, denoted *config* [9], controlled by a user-defined parameter *w* (varying between 0 and 1).

$$Q(w) = compl^{w} * config^{1-w} \qquad (2)$$

Here, *compl = #p / #P* and *config = ((#p / (#p + #n)) – (#P / (#P + #N))) * (#P + #N) / #N.* Here, #p and #n are the numbers of positive and negative examples covered by the rule, and #P and #N are the numbers of positive and negative examples in the training dataset.

In each mode, rules are selected from stars using a *Lexicographical Evaluation Functional* (LEF), a multicriterion measure of rule preference, defined by the user by selecting a subset of elementary criteria from a list of such criteria predefined in AQ21 [18]. Such elementary criteria include, for example, the maximization of the number of positive examples, the number of positive examples not covered by other rules, the rule simplicity, and rule confidence, and the minimization of the cost measuring attributes in the rule. The final hypothesis assembled from stars by selecting the fewest highest LEF-ranked rules that cover all positive examples (in TF mode) or a significant subset of them (in ATF mode).

2.2. Learning Censored Rules

The concept of an exception from a rule is commonly used by humans when they are referring to rarely occurring anomalies in the rule application (e.g., [7], [16]). It is not unusual that a simple rule may work well most of the time but not always. Improving it to work for all cases would make it, however, much more complex. To capture such cases, AQ21, learns *censored rules*, a.k.a. rules with *exception parts,* as presented in (1).

A standard run of AQ21 in all modes creates standard rules, i.e., rules without exception parts. Method of generating exceptions depends on mode in which AQ21 is executed. In ATF and PD modes, where inconsistency is allowed, the system learns standard rules and generates exception phrases that represent covered negative examples. This is done by finding a conjunctive description of negative examples by applying AQ learning to the examples covered by the rule. The result of generating exceptions in ATF mode is a hypothesis that consists of rules that may have exceptions, and in PD mode is a set of patterns with exceptions.

In TF mode, where consistency must be guaranteed, the program adds negative examples to the list of exceptions, if such examples are i*nfrequent* in comparison to the examples covered by the rule, but would introduce significant complexity in order to accommodate them. If all exceptions from the rule can be described by one conjunctive description, such a description is created and used as the exception part, otherwise, an explicit list of examples that are exceptions is used.

2.3. Learning Alternative Hypotheses

From any non-trivial set of concept examples, it is usually possible to generate many alternative inductive generalizations. Such generalizations, called alternative hypotheses, can be useful for a variety of practical applications of computational learning systems. For example, in medical decision making (diagnosis, drug prescription, or therapy assignment), some tests required by a given diagnostic procedure may be unavailable and an alternative procedure would be necessary. Alternative hypotheses can also be used to increase the accuracy of classification decisions, through simple voting on decisions assigned by different hypotheses, or by weighted voting, as is typically done in boosting.

AQ21 learns alternative hypotheses in two steps. In the first step, more than one rule is selected from each star, thus different generalizations of the seed are kept (when k > 1 in the algorithm in Figure 1). In the second step, these rules are assembled together to create alternative hypotheses, ordered based on user-defined criteria. Details of the algorithm for learning alternative hypotheses are presented in [8].

3. Application to Metabolic Syndrome

Metabolic syndrome (MS) and its secondary complications are clinical entities that pose a significant challenge in determining correct diagnosis. Abdominal obesity and insulin resistance appear to be its predominant underlying risk factors. Some types of metabolic abnormalities predispose people to non-alcoholic fatty liver disease (NAFLD) and nonalcoholic steatohepatitis (NASH). The costs associated with treating patients with these complications are substantial, thus an early prediction and prevention of complications is of significant importance. Currently, it is not possible to make an accurate diagnosis without liver biopsy. It is an invasive and costly procedure that is prone to complications, some minor, such as pain, and some more severe, including death [12]. An attractive alternative, pursued in this research, is to use panels of the serum markers, as blood samples could be collected in a minimally invasive way. However, the predictions made in different studies using current non-invasive methods lack consistency. Most of the clinical studies of MS are based on the same serum parameters as ones used in this study, but are performed on only one group of patients collected in a single hospital and use only simple statistical measures for group comparisons and correlation plotting. For example, Poynard et al. developed a panel of biomarkers known as *NashTest* aimed at the detection of NASH in patients with NAFLD in order to reduce the need for liver biopsy [12]. Described panel uses proprietary algorithms taking into account a combination of 13 parameters: age, sex, height, weight, and serum levels of triglycerides, cholesterol, alpha2macroglobulin, apolipoprotein A1, haptoglobin, gamma-glutamyl-transpeptidase, transaminases ALT, AST, and total bilirubin. The specificity (true negatives) of the method is 94% for NASH, and its selectivity (true positives) is 33%.

A similar study presented a composite index for distinguishing steatosis from NASH calculated by counting the risk factors: *age >=50 years*, *female gender*, *AST ≥ 45 IU/l*, *BMI ≥ 30 mg/kg2*, *AAR ≥ 0.80*, and *HA ≥ 55 microg/l*. The presence of three or more of the above risk factors had sensitivity 73.7% and specificity 65.7% [10].

A number of other researchers attempted to use HOMA, or a combination of adiponectin and other secreted adipokines (e.g. resistin, visfatin, leptin etc.) as predictors of NASH. It is well known that low levels of adiponectin expression may predispose patients to the progressive form of NAFLD or NASH [1], [5].

3.1. Data

As mentioned earlier, data used for this study were in aggregated form. They were collected from articles published in journals such as *Hepatology*, *Obesity Research*, *International Journal of Obesity* and some others. The data do not include individual measurements of clinical parameters, as the individual's privacy is protected by HIPAA (Health Insurance Portability and Accountability Act) and similar protection policies.

For this study, we retrieved aggregated clinical data from 16 separate hospital cohorts that included 12 groups of patients with present liver disease symptoms and 8 control groups of healthy subjects. Every single group of patients was described in terms of the mean of attributes measured for this group of patients. The total number of different parameters measured was 152. Different attributes were measured, however, in different studies, which added additional complexity to the problem. The only parameter that was quantified in all studies is AST (level of aspartate aminotransferase).

Twenty most common attributes used in this study are presented in Table 1. In addition, we constructed an extra attribute, defined as the ratio of AST and ALT, as this ratio is often used in liver disease diagnostics.

Table 1: List of attributes used in the initial study.

Attribute Name	**Description**
Weight	Patient's Weight (Kg)
BMI	Body-Mass Index
Height	Patient's Height
Fast glu	Glucose level after fasting (mg/dll)
Fast ins	Insulin level after fasting (mUI/l)
Total cholest	Cholesterol level
HDL	High-Density Lipoproteines (mg/dl)
LDL	Low-density Lipoproteines (mg/dl)
Triglycerides	Triglyceride levels (mg/dl)
HOMA	Level of insulin resistance according to the Homeostasis Model Assessment of insulin resistance
AST	Aspartate aminotransferase level (U/l)
ALT	Alanine aminotransferase level (U/l)
Gamma GT	Gamma-glutamyltransferase level (U/l)
ALP	Alkaline phosphatase level (U/l)
Leptin	Leptin level (ng/ml)
Adiponectin	Adiponectin level (mg/ml)
Systolic bp	Systolic blood pressure
Diastolic bp	Diastolic blood pressure
Body fat %	Percent of the body fat
Visceral fat %	Percent of the body visceral fat

The domain of the output attribute "Class" contains diseases under consideration, namely NAFLD (non-alcoholic fatty liver disease), SS (simple steatosis), and NASH (Nonalcoholic Steatohepatitis), and includes also a healthy status, represented by control groups serving as a contrast set for learning. It should be noted that NAFLD is the most general condition that comprises both SS and NASH cases, which means that values of the output attribute form a hierarchy. Therefore, we first sought rules that differentiate examples of all three diseases together from healthy cases and then rules characterizing NASH, the most severe form of NAFLD.

3.2. Selected Results

In order to explore different types of rules that can be learned from the data, AQ21 was run in PD and TF modes, and with different parameter settings. Selected results that appear be medically most interesting are presented in this section.

In the first set of experiments, we sought rules differentiating the three diseases from the healthy cases provided by control groups. Here is a selection of the censored rules learned by AQ21 working in PD mode. Due to use of exceptions patterns in the first two rules were turn into consistent form (which may not happen in the general case when exceptions cannot be characterized by one attributional statement).

```
[Class=NAFLD or SS or NASH]
   <= [BMI>=26.85: 8,2]
       |_ [AST<=27.2] & [Adiponectin>=7.25]
        : p=8,n=0,q=0.816,cx=25

[Class=NAFLD or SS or NASH]
   <= [HOMA>=2.27: 9,2]
       |_ [Fast_ins<=13.17] & [Leptin>=14.25] &
       [Adiponectin>=7.25]
        : p=9,n=0,q=0.972,cx=35

[Class=NAFLD or SS or NASH]
   <= [Adiponectin<=6.18: 8,1]
         : p=8,nmin=0,nmax=1,q=0.695,cx=5
```

The first rule states:

> *There is presence of non-alcoholic fatty liver disease or its subtypes, simple steatosis or nonalcoholic steatohepatitis, if body-mass index is greater or equal 26.85, except for when aspartate aminotransferase level is at most 27.2 and adiponectin level is at least 7.25.*

A detailed explanation of the parameters in the annotation is in the AQ21 User's Guide [18]. Other rules can be interpreted similarly. In the third rule, the condition part indicates that one negative example is covered. This example is only potentially covered, as its value of adiponectin is unknown, which is indicated as n_{min}=0 and n_{max}=1 in the rule's annotation.

Another set of experiments was to discover rules for diagnosing the most severe form of NAFLD, nonalcoholic steatohepatitis (NASH). Groups of patients with non-alcoholic fatty liver disease (NAFLD) were excluded from dataset for these experiments, as this clinical entity covers both benign steatosis (SS) and aggravated disease (NASH).

The following pattern was learned by AQ21 in PD mode:

```
[Class=NASH]
   <= [AST_ALT_ratio<=0.8121: 4,1]
       |_ [Total_cholest>=209.3]
        : p=4,n=0,q=1,cx=15
```

The rule is clinically relevant as it reflects the most skewed cases of extreme AST/ALT ratios usually pinpointing most severe forms of the non-alcoholic liver pathology.

In AQ21's TF mode, the following complete and consistent alternative rulesets were learned:

Alternative Ruleset 1

```
[Class=NASH]
 <= [Adiponectin<=5.925: 4,1] &
    [AST_ALT_ratio<=9.879: 4,10]
     : p=4,n=0,cx=10
```

Alternative Ruleset 2

```
[Class=NASH]
 <= [AST>=22.88] & [Adiponectin<=5.925]
       : p=4,n=0,cx=10
```

Alternative Ruleset 3

```
[Class=NASH]
 <= [HOMA>=1.86] & [Adiponectin<=5.925]
    : p=3,n=0,cx=10
 <= [Fast_ins>=9.5] & [Adiponectin<=5.925]
    : p=3,n=0,cx=10
```

In the above results, the strongest indication of non-alcoholic fatty liver disease is low value of adiponectin. The condition would be sufficient to describe all groups of patients with nonalcoholic steatohepatitis. However, because of one study in which values of adiponectin were not available, the AQ21 system needed additional conditions to guarantee consistency of the rule in TF mode.

The rules based on the adiponectin concentrations are especially important as the role of the decrease in the adipocytic secretion of this protein is related to an increase of the BMI. Rulesets created by AQ21 unequivocally point to the importance of this secreted protein of adipose as a potential diagnostic component.

6. Summary

This paper briefly reviewed the AQ21 natural induction system and its application to discovering patterns in the aggregated clinical data obtained from different studies of liver diseases associated with metabolic syndrome. AQ21 is particularly attractive for this problem because it seeks hypotheses that are not only accurate but also easy to understand, which is particularly important in medical applications. The system can be viewed as a laboratory for pattern discovery, because is allows the user to experiment with learning of different types of hypotheses (fully consistent, approximate, or patterns), apply different criteria of attributional rule quality, learn hypotheses with and without exception parts, generate an optimized set of alternative hypotheses from the same data, control their level of generality, and some other features [18], [19].

The study produced some clinically relevant results, including one stating the importance of adiponectin, which might be added to the currently used panels of non-invasive diagnostic markers. The learned rules also indicate important relationships between several diagnostic markers and diseases from the metabolic syndrome spectrum. Their further evaluation, on both aggregated and non-aggregated data (not available for us in this study), is needed in order to determine their predictive accuracy.

In our future research, we plan to analyze datasets characterizing other diseases of metabolic syndrome spectrum than those of concern here, to experiment with a variety of parameter settings, and to test the obtained hypotheses on a larger number of cases to ensure high reliability of the predictive accuracy.

References

[1] Baranova, A., Gowder, S.J., Schlauch, K., Elariny, H., Collantes, R., Afendy, A., Ong, J.P., Goodman, Z., Chandhoke, V. and Younossi, Z.M., “Gene expression of leptin, resistin, and adiponectin in the white adipose tissue of obese patients with non-alcoholic fatty liver disease and insulin resistance,” *Obesity Surgery,* 16, 2006, pp. 1118-1125.

[2] Clark, P. and Niblett, T., “The CN2 Induction Algorithm,” *Machine Learning*, 3, 1989.

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

[4] Collantes, R.S., Ong, J.P. and Younossi Z.M. “The metabolic syndrome and nonalcoholic fatty liver disease,” *Panminerva Med*., 48(1), 2006.

[5] Kowdley, K.V. and Pratt, D.S., “Adiponectin--tipping the scales from NAFLD to NASH?,” *Gastroenterology,* 128(2), 2005, pp. 511-513.

[6] Michalski, R.S., “A Theory and Methodology of Inductive Learning,” In R.S. Michalski, J. Carbonell and T.M. Mitchell (Eds.), *Machine Learning: An Artificial Intelligence Approach*, Palo Alto: Tioga Publishing Co. 1983.

[7] 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, 2004.

[8] Michalski, R.S., “Generating Alternative Hypotheses in AQ Learning,” *Reports of the Machine Learning and Inference Laboratory*, MLI 04-6, George Mason University, 2004.

[9] Michalski, R.S. and Kaufman, K., “Learning Patterns in Noisy Data: The AQ Approach,” In G. Paliouras, V. Karkaletsis and C. Spyropoulos, (Eds.), *Machine Learning and its Applications*, Springer-Verlag, 2001, pp. 22-38.

[10] Palekar, N.A., Naus, R., Larson, S.P., Ward, J. and Harrison, S.A. Clinical model for distinguishing nonalcoholic steatohepatitis from simple steatosis in patients with nonalcoholic fatty liver disease. *Liver Int.,* 26(2), 2006, pp. 151-156.

[11] Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning about Data*, Springer, 1991.

[12] Poynard, T., Ratziu V., Charlotte, F., Messous, D., Munteanu, M., Imbert-Bismut, F., Massard, J., Bonyhay, L., Tahiri, M., Thabut, D., Cadranel, J.F., Le Bail, B., de Ledinghen, V., “LIDO Study Group; CYTOL study group. Diagnostic value of iochemical markers (NashTest) for the prediction of non alcoholo steato hepatitis in patients with non-alcoholic fatty liver disease,” *BMC Gastroenterol*., 2006.

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

[14] Setzkorn, C. and Paton, R.C., “On the use of multi-objective evolutionary algorithms for the induction of fuzzy classification rule systems,” *Biosystems*, 81, 2, 2005.

[15] Spinler, S.A., “Challenges associated with metabolic syndrome,” *Pharmacotherapy*, 12, 2006, pp. 209S-17S.

[16] Suzuki, E. and Zytkow, J.M., “Unified algorithm for undirected discovery of exception rules,” *International Journal of Intelligent Systems*, 20, 6, 2005, pp. 673-691.

[17] Van Zyl, J. and Cloete, I., “Simultaneous Concept Learning of Fuzzy Rules,” *Proceedings of the Fifteenth European Conference on Machine Learning*, Pisa, Italy, 2004, pp. 548-559.

[18] Wojtusiak, J., “AQ21 User's Guide,” *Reports of the Machine Learning and Inference Laboratory*, MLI 04-3, George Mason University, Fairfax, VA, 2004.

[19] 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*, 2006.

[20] Zarich, S.W., “Metabolic syndrome, diabetes and cardiovascular events: current controversies and recommendations,” *Minerva Cardioangiol*, 54(2), 2006, pp. 195-214.