Model Learning from Published Aggregated Data

Janusz Wojtusiak and Ancha Baranova

Abstract. In many application domains, particularly in healthcare, an access for individual datapoints is limited, while data aggregated in form of means and standard deviations are widely available. This limitation is a result of many factors, including privacy laws that prevent clinicians and scientists from freely sharing individual patient data, inability to share proprietary business data, and inadequate data collection methods. Consequently, it prevents the use of the traditional machine learning methods for model construction. The problem is especially important if a study involves comparisons of multiple datasets, where each is derived from different open-access publications where data are represented in an aggregated form. This chapter describes the problem of machine learning of models from aggregated data as compared to traditional learning from individual examples. It presents a method of rule induction from such data as well as an application of this method to constructing of the predictive models for diagnosing liver complications of the metabolic syndrome – one of the most common chronic diseases in humans. Other possible applications of the method are also discussed.

1 Introduction

Open – access publications are one of the most important sources of scientific data vital for healthcare research and industry. These publications can be automatically searched and retrieved from the internet and in many instances they are used to build foundation for further studies. Unfortunately, it is often not possible to obtain

Janusz Wojtusiak
Department of Health Administration and Policy,
George Mason University Northeast Module,
Room 108 4400 University Drive, MSN 1J3
Fairfax, VA 22030, USA
e-mail: jwojt@mli.gmu.edu

Ancha Baranova
The Center for Biomedical Genomics
Room 182 Discovery Hall, MSN 4D7
10900 University Blvd
Manassas VA, 20110
e-mail: abaranov@gmu.edu

M. Biba and F. Xhafa (Eds.): Learning Structure and Schemas from Documents, SCI 375, pp. 369–384.
springerlink.com

the original datasets on which the published studies were performed. This is because more often than not, scientists are reluctant or not allowed to share their original data. This is particularly the case in medical, behavioral, and social studies in which data are protected by patients' privacy laws. Many studies also use confidential financial, management or security datasets that cannot be shared outside an organization. The most common reasons for which data are not available are:

- **Patient privacy.** Privacy laws preventing sharing and using individual patient data are enforced in most countries. While particulars may differ, the privacy laws often require informed consent of each patient for his or her personal data to be used for a specific study. Examples of such laws are the U.S. Health Portability and Accountability Act of 1996 [1], and the Directive 95/46/EC of the European Parliament and the Council of 24 October 1995 that cover the protection of individuals with regard to the processing of personal data and the movement of such data [29].

- **Confidentiality.** In many cases collected datasets remain confidential as they include information that cannot be shared due to business or other reasons. This is often the case with financial information (i.e. hospital billing datasets) or existing patients' records (i.e., electronic medical records) that a company keeps protected from competitors. To attain access to such data, special agreements are required, and these are often impossible to arrange.

- **Keep data within a research group for further use.** All aspects of collecting data take enormous efforts. Those who have access to reliable datasets have better chances of publishing their research results and, therefore, better performance reviews and possibility of funding. However, while most researchers do agree that all research data should be shared, not many are actually willing to share their own datasets.

- **Lack of public trust in sharing data.** People are concerned about storage and sharing of their personal information as data by public sector and private organizations. This includes but is not limited to data like DNA fingerprints and electronic medical records.

- **No individual data are recorded.** In many cases no individual data are collected, recorded, or reported in the final dataset(s). Some examples of this type of datasets include public health data that combine reports from multiple local governments or organizations that communicate to public only summaries for each area of assessment, and results of large scale experiments in which datasets are simply too large to store and, therefore, need to be immediately processed and aggregated before storing. Such datasets are found, for instance, in astronomy and physics.

Recently, some research funding agencies, including the National Institutes of Health, implemented a policy requiring that data collected in supported studies, mainly clinical trials, be available to others for research purposes. This policy, however, covers only a small fraction of studies performed worldwide.

Most publications, for many of the above reasons, will include only summaries of data in aggregated forms. This is partially due to public health surveillances and data collection systems, which rely on aggregated data collected by distributed institutions [7]. Traditional machine learning methods are not suited for such datasets, as they are designed to work with individual data points. This chapter focuses on the use of aggregated data extracted from medical publications. However, the methodology is translatable and can be applied in other domains.

2 Mission, Objectives, and Contributions

Analysis of published aggregated data requires a new class of machine learning methods. Aggregated data are most often presented as means and standard deviations for several parameters measured over a group of observations (i.e., in certain cohort of patients). This chapter introduces a concept of a learning process based on aggregated data, and describes a rule-based approach to creating predictive models from such datasets. Specifically, the approach employs an AQ-based learning process that uses aggregated data to derive attributional rules that are more expressive than standard IF … THEN rules. This knowledge representation is briefly described in Section 5, and the learning algorithm in Section 6.

There are several requirements for successful application of machine learning to aggregated data. Many of these criteria are also applicable to traditional machine learning from individual examples.

- **Accuracy.** Models have to provide reliable predictions, which is in most cases their main function. Although models are learned from aggregated data, their accuracy is measured using individual datapoints within a traditional type of validation datasets. Multiple measures of accuracy are available, all of which perform some form of accounting of correct and incorrect predictions and combinations thereof. The most commonly used measures of accuracy include precision, recall, sensitivity, specificity, F-score, and others. When only aggregated data are available, these measures can be estimated as described in Section 6.3.

- **Transparency.** Medical and healthcare studies require models to be understood easily by people not trained in machine learning, statistics, and other advanced data analysis methods. In this sense, providing just the reliable predictions is not sufficient, as models should also "explain" why a specific prediction is made and what the model actually does. The concept of understandability and interpretability is very well known in expert systems and early work on artificial intelligence, but has been largely ignored by many modern machine learning methods.

- **Acceptability.** Models need to be accepted by potential users. While partially related to transparency, acceptability requires that the models don't contradict the knowledge of existing experts or are otherwise "reasonable."

- **Efficiency.** Both model induction and model application algorithms need to be efficient. Although machine learning from aggregated data in many cases does not involve very large datasets, data that is derived from relevant publications can represent between tens and thousands of cohorts. Although much smaller than considered for data mining algorithms, aggregated datasets should not be subjected to analysis by inefficient algorithms with very large computational complexity.

- **Exportability.** Results of machine learning should be directly transferable to decision support systems. It is not unusual that these learned models will work along with other existing models and need to be compatible. For example, learned models can be translated or directly learned in the form of rules in Arden Syntax [14], a popular representation language in clinical decision support systems.

3 Related Work

The problem of analyzing results of published studies is well known. Systematic reviews are used to gather, process, and analyze findings within a collection of closely related studies. Their goal is to arrive at conclusions supported by many other studies. Meta-analysis methods, often used in systematic reviews, are used to calculate statistical descriptions that characterize data used in multiple studies. Systematic reviews and meta-analysis of published studies are important research tools, and are particularly popular in healthcare, policy, social sciences, and law. By combining the results of multiple studies, meta-analysis is able to increase the confidence in study conclusions and cross-validate the results of the particular study. Extensive theory has been built on how to aggregate results from multiple studies and derive statistically valid conclusions [15]. These methodologies are used in preparing systematic reviews such as those by the Cochrane Collaboration [13][4] in healthcare, and the Campbell Collaboration [5][9]**.** in public policy and law.

In addition to other disciplines concerned with identifying knowledge in published studies, although other forms exist including rarely use the same sets of attributes. Literature-based discovery (LBD) seeks to identify unknown relationships in data drawn from published results [12][30]. By bringing together results published in several papers, new relationships that were not considered in the original studies can be found. Several methods have been created to support LBD [3]. While the general framework of LBD is somewhat similar to the described method, its goal is to discover relationships, rather than build models.

Significant work has been done to develop methods for finding and classifying publications to be included in systematic reviews [18]. These include both research and commercial systems working with publication databases.

Surprisingly, despite the extremely fast growth of machine learning, a discipline that developed powerful data analysis and knowledge discovery tools, little work has been done to use advanced learning methods to support systematic reviews and meta-analysis. Two machine learning areas that are closely related to the described method are statistical relational learning [11] [6] and inductive logic programming [16]. Both areas deal with the more general problem of learning from datasets with complicated structures, rather than the specific problem of learning from published results.

4 Aggregated Data

In this chapter we assume a usual situation in which each patient is an individual datapoint described using a set of attributes $A_1 \ldots A_k$. Typical machine learning programs use such individual datapoints in the form of attribute-value examples (1) where $v_1, v_2, \ldots, v_k$ are values of attributes $A_1, \ldots A_k$.

$$(v_1, v_2, \ldots, v_k) \tag{1}$$

Typically each example is described using the same attributes, thus the input dataset used for learning is in the form of a flat attribute-value table. In the case when some attributes are not present in a description of a specific example, meta-values (a.k.a. missing values) can be used. This form of data is, however, almost never included in published manuscripts for reasons outlined in the introduction.

The most typical form of data available in publications is "aggregated tables," although other forms including correlations coefficients, regression models, and others. An aggregated table includes a summary of data aggregated for one or more group of examples (i.e., patient cohorts), usually given as means and standard deviations or frequencies of attributes' values in that groups. Some papers report standard errors, variances, or confidence intervals that can be usually converted into standard deviations. In this chapter we assume that $G_1, \ldots, G_n$ are groups of patients described in publications $P_1, \ldots P_p$. One publication often includes more than one group of patients (i.e., disease and healthy controls, or before and after treatment). Usually, all patients within one group are described using the same set of attributes, although patient groups described in different publications rarely use the same sets of attributes.

Table 1 illustrates example data derived from multiple publications related to metabolic syndrome. It includes means and standard deviations of several attributes in two groups (NAFLD and controls) derived from studies about non-alcoholic fatty liver disease (NAFLD).

Table 1 Example aggregated data derived from multiple publications. It is a subset of datasets used to induce rules described in Section 7.

	NAFLD	NAFLD	SS	NASH	SS	C_NASH
M/F	15–2	155–19	?	?	11–10	10–9
Age	44+/-3	41+/-11	?	?	41+/-13	43+/-14
Weight	86.2+/-3.5	?	?	?	?	?
BMI	27.4+/-0.8	27.3+/-3.2	?	?	33+/-45	31+/-4
Height	1.77+/-0.02	?	?	?	?	?
FG	107.1+/-7.56	98.2+/-26.0	108.90+/-9.09	108.18+/-8.72	93+/-13	91+/-11
FI	13.4+/-1.5	15.1+/-7.9	13.9+/-2.0	12.7+/-2.2	23.98+/-16.78	12.94+/-9.6
TC	208.07+/-11.92	?	211.53+/-19.23	199.61+/-7.69	207+/-36	280+/-50
HDL	48.36+/-3.9	47+/-11	44.85+/-7.8	50.31+/-10.92	36+/-6	47+/-14
LDL	126.36+/-11.15	?	121.68+/-11.7	128.31+/-12.09	135+/-26	131+/-41
T	181.56+/-31.15	138+/-93	191.35+/-40.05	178+/-26.7	230+/-85	165+/-75
HOMA	3.61+/-0.55	?	3.30+/-0.40	3.75+/-0.60	7.0+/-5.4	3.2+/-3.0

FG = Fasting Glucose (mg/dl), FI = Fasting Insulin (mUI/l) , TC = Total Cholesterol, T = Triglycerides

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, each group can be described by an *aggregated example* given as (2). It can be simplified into (3) when order of attributes is defined.

$$(A_1 = (\mu_{A_1}, \sigma_{A_1}), A_2 = (\mu_{A_2}, \sigma_{A_2}), \ldots, A_k = (\mu_{A_k}, \sigma_{A_k})) \tag{2}$$

$$((\mu_{A_1}, \sigma_{A_1}), (\mu_{A_2}, \sigma_{A_2}), \ldots, (\mu_{A_k}, \sigma_{A_k})) \tag{3}$$

For non-numerical attributes, a typically used aggregated form lists frequencies of values in a group, explicitly showing distribution of examples. For example, a group of patients may include 20% smokers and 80% non-smokers.

Another type of data describes entire groups. In medical or social publications these can be related to inclusion criteria for a specific study and additional facts about participants. Although describing groups of data, these attributes refer directly to individual examples. For, example if a study is performed among white males, then each individual subject in the data has precisely this value for attributes describing ethnicity and gender.

Sample sizes (numbers of examples in groups) are always provided. Although they do not provide any information about the subjects themselves, but rather about groups, sample sizes constitute important information crucial during model induction and its coverage estimation. Given both aggregated and not aggregated data, examples take the form (4).

$$(size, (\mu_{A_1}, \sigma_{A_1}),(\mu_{A_2}, \sigma_{A_2}),...,(\mu_{A_k}, \sigma_{A_k})),v_k + 1,...v_l) \tag{4}$$

Note that in one study an attribute can be in the aggregated form, in the second study the same attribute can be in non-aggregated form (i.e. used as inclusion criteria), and completely not available in the third study.

The forms of data outlined above differ from those typically used in machine learning from examples, dominated by learning from individual attribute-value examples in the form (1). Although handling qualitative statements and background knowledge, which is a part of structured machine learning [8], and has been well studied in inductive logic programming [16] [25] and statistical relational learning [11], no special methods for learning from published results that stress aggregated data are available. Furthermore, although relational learning assumes using aggregates [26] [28], they deal only with individual examples with additional characteristics that are being aggregated. This is in contrast to the presented method in which an aggregated example represents a group.

5 Attributional Rules as Knowledge Representation

An earlier part of this chapter discussed requirements for models induced from aggregated data. Rule-based knowledge representation is known to satisfy several of the criteria. However, standard IF…THEN rules are using only conjunctions of simple statements and have limited expression power. Therefore, more expressive forms of rules are used in the presented work.

The main representation of knowledge used in the described method is *attributional rules* [21] whose one form is given by (5). Both *CONSEQUENT* and *PREMISE* are conjunctions of attributional conditions (6). The symbols <=, and |_ denote implication and exception operators, respectively. *EXCEPTION* is either an exception clause in the form of a conjunction of attributional conditions or an explicit list of examples constituting exceptions to the rule. *ANNOTATION* is an additional statistical description, including, for example, the rule's coverage.

$$CONSEQUENT <= PREMISE \mid\_ EXCEPTION : ANNOTATION \tag{5}$$

$$[L\ REL\ R : A] \tag{6}$$

An attributional condition corresponds to a simple natural language statement. Its general form is (6), in which L is an attribute, a counting attribute (derived from other attributes), or a simple arithmetical expression over numerical attributes; R is an attribute value, internal disjunction or conjunction of values, a range, or an attribute; *REL* is a relation applicable to L and R; and A is an optional annotation that provides statistical information characterizing the condition. The annotation includes numbers of cases satisfied by the condition and its consistency. When L is a binary attribute *REL* and R may be omitted. Several other forms of attributional rules are available, all of which resemble statements in natural language, and thus are interpretable by people not trained in machine learning [21].

The above choice of rule-based knowledge representation is based on the fact that it satisfies transparency and exportability criteria for models stated in Section 2. It can also provide accuracy comparable with other representations, without the need to employ special procedures that convert black-box representations to human-oriented explanations [2].

The following section describes an algorithm for inducing attributional rules from data, and its extension needed to handle aggregated data.

6 Rule Induction

6.1 AQ Algorithm

Many algorithms are available for inducing rules from data. Despite their differences, the algorithms have two common elements: rule construction, and rule evaluation. Although the described method for learning from aggregated data can be adapted to most rule learning systems, in this chapter the focus is on the AQ approach to rule learning [19] [20] [32]. This focus is important because the method has several advantages that make it suitable for learning from aggregated data. AQ generates attributional rules described above, deals with multiple data types [23] and meta-values [22], includes different generalization and reasoning methods, and is fairly flexible due to the large number of parameters that control the learning process. The method follows the popular separate-and-conquer approach to rule learning that is summarized by [10], and is capable of using powerful statistical measures of rules quality that incorporate aggregated data found in published results.

The 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. Within the star generation, AQ applies an *extension-against operator[20]* whose goal is to find all possible rules that distinguish a given positive example, called *seed*, from a given negative example. In the original method, the extension-against is a purely logical operation and it is denoted by the --| symbol. In its simplified form for non-aggregated data, a seed $s = (a_1,\ldots a_k)$ extended against a negative example $n = (b_1,\ldots b_k)$ is a set of one-condition rules shown in (7) for all attributes for which values in the seed and the negative example are different.

$$s \text{ --| } n = \bigvee [A_i \neq b_i] \text{ for all } i \text{ such that } a_i \neq b_i \tag{7}$$

For example, (7, 5, 3) --| (2, 5, 4) = $[A_1 \neq 2] \vee [A_3 \neq 4]$. Here, the attribute A_2 is not used because it takes the same value in the seed and the negative example. The operator works the same way for symbolic (nominal, structured, etc.) and numeric (interval, ratio, etc.) attributes.

Multiple applications of the operator allow for the creation of rules that cover the seed and exclude negative examples. Intersection of all such rules covers the seed and rule out any negative example.

At this stage of rule construction, AQ applies a beam search to filter out potentially large number of generated rules. The method allows for multiple criteria of rule evaluation, most of which are based on statistical evaluation and complexity of rules. In the second stage, rules/hypotheses are optimized to maximize their predictive accuracy while maintaining simplicity. This process is somewhat similar to pruning, which is frequently done by learning programs. In AQ, rules are not only pruned, but can also be extended through a set of optimization operators working on attribute, condition, rule, and hypothesis levels.

When learning from aggregated data, information about distributions of values in aggregated examples is used. Consequently, the *extension-against operator* is no longer purely logical.

6.2 Rule Induction from Aggregated Data

A general schema of learning from published results is presented in Figure 1. Aggregated and individual data are used for rule generation and evaluation, while qualitative/quantitative results and background knowledge are used to constrain the generation of models. Each rule is evaluated not only for coverage-based quality (statistical measures), but also by its simplicity and given constraints. There are several possible approaches to the problem of rule induction from aggregated examples. This section briefly overviews these approaches, with the focus primarily on the third method that directly uses aggregated information within the AQ rule induction algorithm.

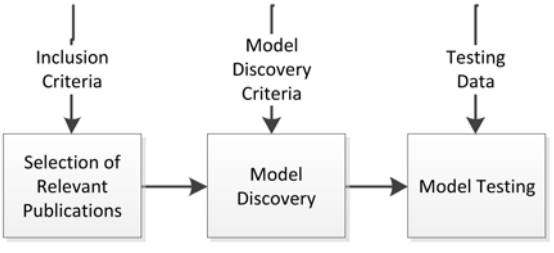

Fig. 1 A flowchart describing the process of model development based on published aggregated data.

Sampling. One simple approach for learning from published results is to approximate the original datasets by sampling. An initial study [24], in which aggregated examples were sampled, indicated, however, that the method does not work well due to deficient information of interrelationships between attributes. An important advantage of the method is that any machine learning method from examples can be applied, because individual examples are created during the sampling process.

Weighting. Another simple method is to use aggregated examples in the form (8) which includes only mean values. Standard deviations are used to weight examples [33] [34].

$$(\mu A1, \mu A2, \ldots, \mu Ak) \tag{8}$$

By doing so, there is no need to significantly modify rule learning algorithms. However, the method ignores important information of "overlaping" groups.

It is important to note that when using this method the rule induction algorithm treats aggregated examples as individual subjects and de facto learns rules that describe weighted groups. Despite its simplicity, the method gave good results in an initial study presented in Section 7.

Limitations of the above methods show the importance of using an algorithm that directly induces rules for classifying individual examples based on aggregated examples. Such algorithm needs to effectively use information about distributions and "recognize" the fact that it is dealing with aggregated examples representing groups not with individual subjects.

Extension-against. The AQ rule learning algorithm can use directly the form (4) of aggregated examples, and effectively incorporate the information about standard deviations when comparing aggregated examples. Assuming normal distribution, $N(\mu_{A_i}, \sigma^2_{A_i})$, over 95% of data described by the aggregated examples lay in the range given by condition (9).

$$[\mu_{A_i} - 2\sigma_{A_i}, \mu_{A_i} + 2\sigma_{A_i}] \tag{9}$$

Thus, when comparing two aggregated examples, a reasonable assumption is that two aggregated values are indistinguishable if the ranges (9) in the examples are intersecting. The modified extension against operator is defined by the formula (10) for numeric attributes.

$$s - -| n = \bigvee [x_i \neq (\mu_{A_i{}^n} - 2\sigma_{A_i{}^n}\ \mu_{A_i{}^n} + 2\ \sigma_{A_i{}^n})] \tag{10}$$

for all i=1..k such that $[\mu_{A_i{}^s} - 2\sigma_{A_i{}^s}\ \mu_{A_i{}^s} + 2\ \sigma_{A_i{}^s}] \cap [\mu_{A_i{}^s} - 2\sigma_{A_i{}^s}\ \mu_{A_i{}^s} + 2\ \sigma_{A_i{}^s}]$.

For aggregated discrete attributes the extension-against operator is defined by the formula (11).

$$s \text{ --| } n = \bigvee [A_i \neq v_1..v_k],\ i = 1..n \tag{11}$$
$$fs(A_i, v_j) < fn(A_i, v_j) + \varepsilon$$

Here, *fs* denotes distribution of values in *s* and *fn* denotes distribution of examples in *n*. For example, if $D(A_1)=\{a,b,c,d\}$. $D(A_2)=\{r,g\}$, $s=(A_1=(0.3,0.05,0,0.65), A_2=(0.2,0.8))$ and $n=(A_1=(0.5,0,0.3,0.2), A_2=(0.27,0.73))$, and ε=0.1, then s --| n = $[A_1 \neq a, c]$. The attribute A_2 is not used at all, because the difference between distributions for that attribute is within the margin ε.

For non-aggregated attributes, i.e. attributes that describe entire groups, the extension-against operator is not modified.

6.3 Calculating Coverage

A key component to any rule induction algorithm is the calculation of rules' coverage. Positive and negative coverage needs to be estimated for individual examples using aggregated examples representing groups. This is needed during both rule creation and rule optimization. In order to estimate numbers of examples from a group satisfying a condition [A=a...b] learned from aggregated data (A is a continuous attribute), the probability (12) of an individual example satisfying the condition can be multiplied by the number of examples in the group. $\Phi_{\mu\sigma^2}(A)$ is the cumulative distribution function, μ is mean value of A in the group, and σ is standard deviation of A in the group.

$$p(A = a...b) = \Phi_{\mu,\sigma^2}(b) - \Phi_{\mu,\sigma^2}(a) \tag{12}$$

In order to estimate the numbers of examples satisfying a rule an independence of attributes is assumed. The rationale behind the assumption is that if two attributes were dependent, the rule learning program would not need to use both of them in the rule (as the value of one implies the value of another). Thus, probabilities (12) for all conditions in the PREMISE can be multiplied. The resulting joint probability is then multiplied by one minus the joint probability of the EXCEPTION, which gives the probability of an example in the group satisfying the rule. Finally, the estimated number of examples from the group satisfying the rule is calculated by multiplying the joint probability by the number of examples in that group. The operation is repeated for all groups for which aggregated data are available.

In the presence of additional information such as covariance between attributes, it is possible to calculate a better estimation of the joint probability than when assuming independence of conditions.

Using learned rules to classify new examples is straightforward, because they are intended to classify individual examples, not aggregated examples representing groups. Rules for classifying individual examples are learned from aggregated examples.

7 Evaluation

The initial methodology [33] [34] for learning from aggregated data has been applied to a small database derived from clinical research publications in a number of well-known peer-reviewed journals. Part of the database was presented in Table 1. The application was concerned with creating predictive models for diagnosing liver complications in metabolic syndrome (MS). Metabolic syndrome and its secondary complications pose a significant challenge for practicing diagnosticians. Abdominal obesity appears to be its predominant underlying risk factor. Metabolic

abnormalities associated with MS, particularly a resistance to the insulin, predispose people to non-alcoholic fatty liver disease (NAFLD) and its more severe manifestation, nonalcoholic steatohepatitis (NASH). The health-related costs associated with these complications are substantial, thus early prediction and prevention of these conditions are of significant importance. Currently, it is not possible to make an accurate diagnosis of NAFLD and/or NASH without a liver biopsy. It is an invasive and costly procedure that is prone to complications, some minor, such as pain, and some more severe, including possibility of death as a result of bleeding or infection [27].

An attractive alternative, pursued in the research, was to use panels of the serum markers, because blood samples could be collected in a minimally invasive way. However, the predictions made in prior studies using currently available prediction algorithms lack consistency. Typically, clinical studies of MS and its complications are performed on single groups of patients collected in one hospital, and use only simple statistical measures for group comparisons and correlation plotting. No large datasets concerning metabolic syndrome are available and data describing measurables in each patients are not available, either. Thus, only methods that deal with results published in papers are applicable.

The data used for this study were in the aggregated form (3). They were collected from articles published in peer-reviewed journals including Hepatology, Obesity Research, International Journal of Obesity, and some others. For the pilot study, we retrieved aggregated clinical data from 20 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 attributes retrieved from papers was 152. In each study however, different attributes were measured, which added additional complexity to the problem. In fact, none of the attributes were present in all studies, even though these very similar studies were dealing with exactly the same clinical problem.

The goal was to construct a set of rules for predicting non-alcoholic fatty liver disease (NAFLD), simple steatosis (SS), and Nonalcoholic Steatohepatitis (NASH). Data also included a number of healthy cohorts, represented as 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. Therefore, we first sought rules that differentiate NAFLD from healthy cases and then rules characterizing NASH, the most severe form of NAFLD.

Below we present two example rules derived by the method. The first rule states that there is presence of non-alcoholic fatty liver disease or its subtypes, if body-mass index is greater or equal 26.85, except for when aspartate aminotransferase level is at most 27.2 units/L and adiponectin level is at least 7.25 mg/ml. The rule's condition is satisfied by eight groups of patients belonging to NAFLD or its subtypes, and two control groups. The exception part of the rule which consists of two conditions filters out both control groups. The entire rule is satisfied by eight groups of patients belonging to NAFLD or its subtypes and non-control groups. The rule's quality is 0.816, and its complexity is 25. The second rule can be interpreted in an analogous way.

```
[Class=NAFLD]
      <== [BMI>=26.85: 8,2]
        |_ [AST<=27.2] & [Adiponectin>=7.25]
         : p=8,n=0,Q(w)=0.816,cx=25
```

```
[Class=NAFLD]
  <== [Adiponectin<=6.18: 8,1]
   : p=8,n_min=0,n_max=1,Q(w)=0.695,cx=5
```

Similar rules have been obtained for predicting simple setatosis and nonalcoholic steatohepatitis [34]. The rules are easy to interpret and are consistent with experts' existing knowledge. An explanation of the parameters is in the AQ21 User's Guide [31], the system that was used to implement the initial methodology.

Validation of these rules for predicting NAFLD resulted in a positive predictive value (PPV) of 85-87%, reflecting relatively high "rule-in" characteristic of the algorithm. The best rule for the prediction of NASH relied on combination of fasting insulin, HOMA and adiponectin values with an accuracy of 78%, with PPV of 71% and negative predictive value (NPV) of 37%.

The models generated by AQ21 are presented in the form of attributional rules, a highly transparent representation which is easy to understand by people not trained in advanced statistics, machine learning, and other computational technologies. Additionally, these kinds of models could be readily imported into existing clinical decision support systems and useful in the settings of point-of-care (POC) initial health assessment. Simplicity of the developed models allows for the use of them on "the back of envelope" in settings where advanced diagnostics are not available (i.e. in developing countries).

8 Discussion

The presented methodology for learning from aggregated data has been developed within the well known AQ rule induction algorithm. The algorithm is able to induce from aggregated data attributional rules for classifying individual examples. The process is depicted in Figure 2.

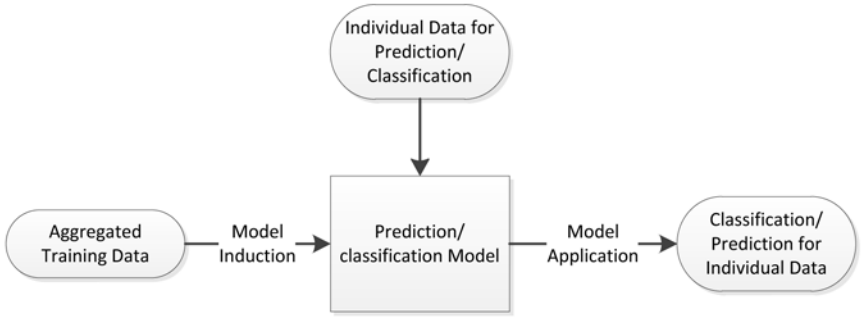

Fig. 2 Induction of models from aggregated data and application to individual data.

Traditional machine learning from examples methods are able to deal with aggregated data when using the sampling method described in Section 6. It is also possible to extend some the methods to deal directly with aggregated data, however, such extension depends on how the specific algorithm works and how it treats individual examples.

Data used in the presented study were manually retrieved from selected publications. It's been recommended that this very labor-intensive process be performed by at least two independently working people and then the results compared and discrepancies discussed in a panel. The selection process of publications to be included in the analysis should also be done by the panel [17]. With the use of currently available technology the process cannot be fully automated, because it depends on the understanding of the publications. However, it is possible to aid personnel in performing this time consuming task. Relevant publications can be pre-selected using advanced search tools available for databases such as PubMed. Data tables can be automatically identified in publications and derived in a tabular form. Ontologies and dictionaries can be used to discover discrepancies in terminology and units. Finally, text mining methods can be used to identify and retrieve data not present in tabular forms (i.e. inclusion criteria).

9 Conclusion

New methods are needed to create accurate and transparent predictive models from de-individualized published clinical data in aggregated forms. Machine learning of attributional rules from published data, including aggregated clinical parameters, inclusion criteria, demographic information and target diagnoses, is able to derive such models.

This chapter described a methodology for machine learning or attributional rules from aggregated published data. The described methodology may complement current systematic reviews and meta-analyses such as Cochrane Reviews. With rapidly changing clinical knowledge, such an automated method with the ability to incrementally update knowledge can prove to be the needed method to keep reviews up to date.

Preliminary application of an early implementation of the method resulted in a set of attributional rules for predicting non-alcoholic fatty liver disease and its subtypes in patients with metabolic syndrome. It illustrated validity of the method on a real-world important problem. Machine learning software applied to the meta-analysis of the published data may provide an easy, non-invasive way to diagnose most patients with NAFLD and NASH. Clinical parameters highlighted by machine learning process can be combined with other non-invasive biomarkers for NASH to increase their accuracy and test characteristics.

References

[1] Annas, G.J.: HIPAA Regulations — A New Era of Medical-Record Privacy? New England Journal of Medicine 348, 1486–1490 (2003)

[2] Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.: How to Explain Individual Classification Decisions. Journal of Machine Learning Research 11, 1803–1831 (2010)

[3] Burza, P., Weeber, M.: Literature-based Discovery. Springer, Heidelberg (2008)

[4] The Cochrane Collaboration, The Cochrane Manual 4 (2008) (updated August 14, 2008)

[5] Davies, F., Boruch, R.: The Campbell Collaboration Does for Public Policy what Cochrane Does for Health. BMJ 323, 294–295 (2001)

[6] De Raedt, L.: Logical and Relational Learning. Springer, Heidelberg (2008)

[7] Diamond, C.C., Mostashari, F., Shirky, C.: Collecting And Sharing Data For Population Health: A New Paradigm. Health Affairs 28(2) (2009)

[8] Dietterich, T.G., Domingos, P., Getoor, L., Muggleton, S., Tadepalli, P.: Structured machine learning: the next ten years. Machine Learning 73(1), 3–23 (2008)

[9] Farrington, D.P., Petrosino, A.: The Campbell Collaboration Crime and Justice Group. Annals of the American Academy of Political and Social Science 578, 35–49 (2001)

[10] Fürnkranz, J.: Separate-and-conquer rule learning. Artificial Intelligence Review 13, 3–54 (1999)

[11] Getoor, L., Taskar, B. (eds.): Introduction to statistical relational learning. MIT Press, Cambridge (2007)

[12] Gordon, M., Lindsay, R.K., Fan, W.: Literature-Based Discovery on the World Wide Web. ACM Transactions on Internet Technology 2(4), 261–275 (2002)

[13] Higgins, J.P.T., Green, S. (eds.): Cochrane Handbook for Systematic Reviews of Interventions (2008), `http://www.cochrane-handbook.org` Version 5.0.0 (updated February 2008)

[14] Hripcsak, G.: Writing Arden Syntax medical logic modules. Computers in Biol-ogy and Medicine 24(5), 331–363 (1994)

[15] Hunter, J.E., Schmidt, F.L.: Methods of Meta-Analysis, Correcting Error and Bias in Research Findings, 2nd edn. Sage Publications Inc., Thousand Oaks (2004)

[16] Lavrac, N., Dzeroski, S.: Inductive Logic Programming: Techniques and Applications. Ellis Horwood, New York (1994)

[17] Lipsey, M.W., Wilson, D.: Practical Meta-Analysis. Sage Publications, Thousand Oaks (2000)

[18] Matwin, S., Kouznetsov, A., Inkpen, D., Frunza, O., O'Blenis, P.: A new algorithm for reducing the workload of experts in performing systematic re-views. Journal of the American Medical Informatics Association 17(4), 446–453 (2010)

[19] Michalski, R.S.: On the Quasi-Minimal Solution of the General Covering Prob-lem. In: Bled, Y. (ed.) Proceedings of the V International Symposium on Information Processing (FCIP 1969), vol. 3, pp. 125–128 (1969)

[20] Michalski, R.S.: A Theory and Methodology of Inductive Learning. In: Michalski, R.S., Carbonell, T.J., Mitchell, T.M. (eds.) Machine Learning: An Artificial Intelligence Approach, pp. 83–134. TIOGA Publishing Co, Palo Alto (1983)

[21] 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 (2004)
[22] Michalski, R.S., Wojtusiak, J.: Reasoning with Missing, Not-applicable and Irrelevant Meta-values in Concept Learning and Pattern Discovery, Technical Report 2005-02, Collaborative Research Center 637, University of Bremen, Germany (2005)
[23] Michalski, R.S., 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)
[24] Michalski, R.S., Wojtusiak, J.: The Distribution Approximation Approach to Learning from Aggregated Data, Reports of the Machine Learning and Inference Laboratory, MLI 08-2, George Mason University. Fairfax, VA (2008)
[25] Muggleton, S.H., De Raedt, L.: Inductive logic programming: Theory and me-thods. Journal of Logic Programming 19(20), 629–679 (1994)
[26] Perlich, C., Provost, F.: Distribution-based aggregation for relational learning with identifier attributes. Machine Learning 62, 65–105 (2006)
[27] 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 biochemical markers (NashTest) for the prediction of non alcoholo steato hepatitis in patients with non-alcoholic fatty liver disease. BMC Gastroenterology 6(34) (2006)
[28] Vens, C.: Complex aggregates in relational learning. AI Communications 21, 219–220 (2008)
[29] Verschuuren, M., Badeyan, G., Carnicero, J., Gissler, M., Asciak, R.P., Sakkeus, L., Stenbeck, M., Devillé, W.: and For The Work Group on Confidentiality and Data Protection of the Network of Competent Authorities of the Health Information and Knowledge Strand of the EU Public Health Programme (August 2003) ; The European data protection legislation and its consequences for public health monitoring: a plea for action. European Journal of Public Health 18(6), 550–551 (2008) doi:10.1093/eurpub/ckn014
[30] Weeber, M., Kors, J.A., Mons, B.: Online tools to support literature-based discov-ery in the life sciences. Briefings in Bioinformatics 6(3), 277–286 (2005)
[31] Wojtusiak, J.: AQ21 User's Guide, Reports of the Machine Learning and Infe-rence Laboratory, MLI 04-3, George Mason University. Fairfax, VA (2004)
[32] Wojtusiak, J., Michalski, R.S., Kaufman, K., Pietrzykowski, J.: The AQ21 Natural Induction Program for Pattern Discovery: Initial Version and its Novel Features. In: Proceedings of The 18th IEEE International Conference on Tools with Artificial Intelligence, Washington D.C (2006)
[33] Wojtusiak, J., Michalski, R.S., Simanivanh, T., Baranova, A.V.: The Natural Induction System AQ21 and Its Application to Data Describing Patients with Metabolic Syndrome: Initial Results. In: Proceedings of the International Conference on Machine Learning and Applications, Cincinnati, OH (2007)
[34] Wojtusiak, J., Michalski, R.S., Simanivanh, T., 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 In-formatics 78(12), e104–e111(2009)