# Applications of machine learning: a medical follow up study

**Du Junping† K. L. Rasmussen‡ J. Aagaard‡ Brian H. Mayoh† Tom Sørensen†**

†Computer Science Department, Aarhus University, Denmark
Email: junping@daimi.aau.dk brian@daimi.aau.dk
‡ Department of Obstetrics and Gynaecology, Aarhus University Hospital, Denmark

## Abstract

This paper describes preliminary work that aims to apply some learning strategies to a medical follow-up study. An investigation of the application of three machine learning algorithms-1R, FOIL and InductH to identify risk factors that govern the colposuspension cure rate has been made. The goal of this study is to induce a generalised description or explanation of the classification attribute, colposuspension cure rate (completely cured, improved, unchanged and worse) from the 767 examples in the questionnaires. We looked for a set of rules that described which risk factors result in differences of cure rate. The results were encouraging, and indicate that machine learning can play a useful role in large scale medical problem solving.

## 1 Introduction

One of the central problems of the information age is dealing with the enormous amount of raw information that is available. More and more data is being collected and stored in databases or spreadsheets. As the volume increases, the gap between generating and collecting the data and actually being able to understand it is widening. In order to bridge this knowledge gap a variety of techniques known as data mining or knowledge discovery is being developed. Knowledge discovery can be defined as the extraction of implicit, previously unknown, and potentially useful information from data, and can be built upon a variety of technologies of which machine learning is one of the most important [Fayyard et al., 1996; Piatetsky-Shapiro and Frawley, 1991].

Machine learning is a technique that can discover previously unknown regularities and trends from diverse datasets, in the hope that machines can help in the often tedious and error-prone process of acquiring knowledge from empirical data, and help people to explain and codify their knowledge and expertise. It encompasses a wide variety of techniques used for the discovery of rules, patterns and relationships in sets of data and produces a generalisation of these relationships that can be used to interpret new, unseen data [Michie, 1991; Pazzani and Kibler, 1992].

WEKA, the Waikato environment for knowledge analysis, is an experimental software workbench incorporating several standard machine learning techniques [McQueen et al., 1994; Witten et al., 1993]. One of the most important features of the WEKA workbench is that it allows many different schemes to be run on the same dataset and for the output of each scheme to be evaluated in a consistent fashion [Holmes et al., 1994]. The WEKA machine learning workbench is used to produce rules and decision trees based on the current dataset. With it, we are able to derive knowledge from datasets that are far too large to be analysed by hand.

At present most research effort in machine learning is directed towards the invention of new algorithms for learning and much less into gaining experience in applying them to important practical applications. Our study explores what machine learning can do in the medical domain. Colposuspension is widely accepted as the best form of suspension in the treatment of female urinary stress incontinence. Numerous risk factors influence cure rate, among them age, body mass, the number of previous operations and the number of deliveries of patients. The aim of our study is to infer the rules that help doctors to identify and recognise the effect of these risk factors on the long term subjective cure rates, that is completely cured, improved, unchanged and worse of patients who underwent a colposuspension procedure in the period 1974 to 1992. These rules can also be used to predict the colposuspension cure rate on the basis of risk factors and could be embedded in automatic processes such as expert systems, or used directly for medical decision making purpose.

## 2 Material and Method

### 2.1 Dataset Preparation

A total of 960 persons who were undergoing colposuspension between August 1974 and December 1992 entered the study. The patients were admitted to three gynaecological departments in Aarhus, Denmark. All records were reviewed; and data from the operation, previous operation, medical history and physical examination were noted. A letter describing the purpose of the study and a detailed questionnaire was sent to the patient. There were two datasets available. Dataset1 was

filled out by the doctors; dataset2 was filled out by the patients. At the time the questionnaire was mailed, 78 of the patients had died. Of the 882 patients who contacted 111 were lost for follow up, so dataset1 contained more data records than that of dataset2. As the identification number (ID) of a patient was the only clue to identify a patient, we compared the ID's of each patient in these two datasets to produce a combined dataset3 (using EXCEL) that contained all the necessary and valid attributes needed for our investigation. There were 767 instances left in dataset3 that was used in our study.

### 2.2 Data Pre-processing and Attributes Selection

With most real world data a significant amount of pre-processing is necessary before the data can be presented to a machine learning system. Typical pre-processing includes the cleaning of noisy, anomalous or missing data. The attribute selection includes variables present in the raw dataset and derived attributes generated from existing variables.

In our experiment, patients were classified on the cure rate attribute, which takes the values: cured, improved, unchanged and worse. In the questionnaire ID number of patient and the date of operation were used, but an absolute date of the operation was not meaningful. The age of patient during operation would be useful, but it was not explicitly present in the dataset. In EXCEL, we used mathematical formulas to get the birthday of a patient from her ID number, after comparing it with the date of the operation, we got the actual age of a patient at the age of the operation. In the dataset only the height and weight of patient were present. As body mass is strongly associated with the cure rate, we used mathematical division in EXCEL to get the body mass of the patient at the time of the operation. In discussions with the doctor at the hospital, it was suggested that the number of previous operations and the number of deliveries were also strongly associated with the cure rate, so they were all selected as attributes. The final attributes selected were age, body mass, the number of previous operations and the number of deliveries of patients.

### 2.3 Dataset Format

WEKA stores data in a common file format called ARFF (attribute relation file format), presenting users with a consistent view of the data regardless of the machine learning scheme being used. ARFF defines a dataset in terms of a relation made up of attributes of data. Information about the names of the relation and the types of the attributes is stored in a header, with the instances being represented as rows of data in the body of the file. Converting the dataset to an ARFF includes loading original data onto a Macintosh spreadsheet package (EXCEL), saving the file in comma separated form, then loading it into our UNIX machine with WEKA installed.

### 2.4 The Machine Learning Schemes

The machine learning tools used for our analysis were primarily FOIL [Quinlan, 1990; Quinlan, 1991; Quinlan, 1993], InductH [McQueen et al., 1994]. and 1R [Holte, 1993]. These are supervised learning schemes that produce useful rules that describe a classification based on combinations of attribute tests. Supervised learning is when a desired class is assigned to each example in the dataset, and the aim is to induce rules that classify unseen examples. By running the dataset3 through each of the schemes individually, the machine learning workbench was used to produce rules about the effects of risk factors on the cure rate.

### 2.5 Output Processing

The output rule is converted into an internal WEKA rule format and evaluated. The rule format is PROLOG-based, and a rule can be executed using an evaluator called PREval. PREval takes a set of rules and an ARFF file, and evaluates how well the rules cover the classifications. It provides figures for classification accuracy, including the percentage correctly classified, incorrectly classified, classified by multiple rules, and not classified at all.

## 3 Results

### 3.1 1R Results

The 1R algorithm for machine learning is a very simple one that proves surprisingly effective. It produces simple rules that choose just one attribute as the criterion for the current decision being made. This scheme will generate a rule stating which attributes are the most effective for deciding the operation result. While Holte uses 1R as a stand-alone learning scheme, we view it as a feature selector. Applying 1R iteratively with each of the attributes in the raw data allows us to rank attributes by their classificatory power.

The rules generated by 1R are listed from best to worst, with the best being listed first. For our dataset 1R has predicted body mass as being the highest ranking at 80.9% accuracy. Age has 67.5% accuracy; the number of previous operations and the number of deliveries have 67.1% accuracy. The most obvious result, and the one we expected, was that the younger the patient, the better the operation result. An example rule produced by 1R looks like this:

Rule for 'Age':

'Cure_rate' ('A') : — 'Age' (X), X <31. % 8/13

'Cure_rate' ('B') : — 'Age' (X), 31 =< X. % 508/752

1Rw accuracy 67.5% (516/765).

The above rule can be explained as this: If age of patient is younger than 31 then the class attribute "Cure_rate" is "completely cured". Otherwise if age of patient is older than or equal to 31 then the class attribute "Cure_rate" is "improved". The Age attribute was accurate 67.5% of the time; 516 patients were correctly classified, 249 patients were unclassified or incorrectly classified and 2 patients did not have a well-defined Age attribute.

### 3.2 InductH Results

InductH produces either independent rules or tree-like rules, where at least one of the results will classify an instance. For our dataset with InductH evaluation, 337 instances were correctly classified, with an accuracy rate of 43.94%; 22 instances were incorrectly classified, with an error rate of 2.86%. One rule produced by InductH looks like this:

'Cure_rate' = 'A' : IF 'No_operation' >= 1.5 [206/584] AND 'No_delivery' < 4.5 [197/545] AND 'Age' < 69.5 [195/537] AND 'No_operation' <5.5 [188/512].

The above rule can be interpreted as this: There are 584 instances where the number of previous operations is larger than 1.5, but only 206 of them have the class attribute "Cure_rate" value of "completely cured". There are 545 instances where the number of deliveries is less than 4.5, but only 197 of them have the class attribute "Cure_rate" value of "completely cured" and satisfy the condition that the number of previous operations is larger than 1.5. There are 537 instances where the age of patient is less than 69.5, but only 195 of them have the class attribute "Cure_rate" value of "completely cured" and satisfy the condition that the number of previous operations is larger than 1.5 and the number of deliveries is less than 4.5. There are 512 instances where the number of previous operations is less than 5.5, but only 188 of them have the class attribute "Cure_rate" value of "completely cured" and satisfy the condition that the number of previous operations is larger than 1.5, the number of deliveries is less than 4.5 and age of patient is less than 69.5.

### 3.3 FOIL Result

FOIL (first order inductive learner), induces logical definitions, expressed as Horn clauses, from data presented in the form of relations. An important feature of FOIL is its ability to express relationships between the attributes in an example. It begins with a set of relations; each defined as a set of related values. Given a particular target relation, it attempts to find clauses that define that relation in terms of itself and other relations. This approach leads to more general, functional definitions that might be applied to new objects.

For our dataset with FOIL evaluation, 676 instances were correctly classified, with an accuracy rate of 88.14%; no instances were incorrectly classified, with an error rate of 0.00%. The rule produced by FOIL looks like this:

'B' (Age, No_operation, No_delivery, body_mass) : — Age > 60, No_delivery> 1, No_operation <= 1, body_mass <= 3

Here 'B' stands for class attribute "Cure_rate" that takes the value "improved".

The above rule can be explained as this: The class attribute is "improved" provided that age of patient is older than 60, the number of deliveries is between 1 and 3 and the number of previous operations is less than or equal to 1.

### 3.4 Evaluation

One of the most important features of the WEKA Workbench is that it allows many different schemes to be run on the same set of data and for the output of each scheme to be evaluated in a consistent fashion. The rules produced by a learning scheme are translated into an equivalent Prolog representation and evaluated with respect to the training and test data sets. For each rule the evaluator indicates how often the rule is used and how many examples were classified correctly and incorrectly. Many schemes have only minimal internal evaluation methods so the Prolog evaluator is very useful for analysing the output from these schemes. It is also a valuable tool for evaluating the performance of different schemes on neutral ground.

The experimental editor in WEKA can be used to evaluate the results of these schemes using a variety of testing methods. In our study, we use hold-out testing method to estimate the accuracy of the rules learned by the algorithms. We specify the ratio of the size of test and training sets and the number of runs required before starting the experimental editor. Results are stored in a text file and processed to provide summary statistics from the PREval evaluator. In our experiment we get an average classification accuracy of 68.97% for 1R; 50.96% for InductH and 86.59% average multiple classification accuracy for FOIL.

## 4 Discussion

The rules generated by 1R were simple in structure, because it only selects one attribute as the criterion at a time. For attribute "Age", it gave a very precise statement about the effect of age on class attribute. Although attribute "body_mass" obtained highest ranking of 80.9% accuracy, which means 619 out of 765 instances were correctly classified, the 1R algorithm did not give a very accurate statement about the effect of body mass on the cure rate.

The rules produced by FOIL are more complex in structure than either 1R and InductH rules. FOIL achieved an accuracy rate of 88.14%, which means 676 of 765 instances were correctly classified. This was the highest accuracy rate obtained among these three algorithms. FOIL expressed the combination effect of different attributes on the class attribute "Cure_rate".

Unfortunately InductH gave relatively low accuracy rate of 43.94% in our study. This may suggest more attributes need to be added after discussions with doctors in the hospital, in the hope that good results can be generated with InductH in the future.

## 5 Conclusion

Machine learning is a burgeoning new technology with a wide range of potential applications. At present, most research effort is directed towards the invention of new algorithms for learning and much less into gaining experience in applying them to real problems. Our paper

redresses this imbalance by grounding machine learning techniques in important practical application. In machine learning, domain knowledge is necessary to analyse data effectively, and our discussions with medical experts direct data processing, experimentation, and interpretation of results. According to the doctors, the following extracted rules are clearly meaningful.

- If a patient was older, her operation result may get worse.
- The more the number of deliveries of patients, the worse the operation result will be.
- The more the number of previous operations of patients, the better the cure rate will be. However the cure rate will fall if the number of previous operations exceeds a certain limit.

The results obtained from our study are encouraging; They indicate that machine learning can play a useful role in large-scale medical problem solving.

## References


[Fayyard et al., 1996] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy.G. *Advances in Knowledge discovery and data mining* . AAAI Press / The MIT Press, Menlo Park, CA. 1996.

[Gaines, 1991] Gaines, B.R. The trade-off between knowledge and data in knowledge acquisition. In Piatetsky-Shapiro and Frawley. AAAI Press, 1991.

[Holmes et al., 1994] Holmes, G., Donkin, A., and Witten, I. H. Weka: A machine learning workbench. *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems,* pages 357-361, Brisbane, Australia, 1994.

[Holte, 1993] Holte, R.C. Very simple classification rules perform well on most commonly-used datasets. *Machine Learning*. 1993; 11: 63-91.

[McQueen et al., 1994] McQueen, R.J., Neal, D., De War, R. and Garner. Preparing and processing relational data through the WEKA machine learning workbench. Working paper, Department of Computer Science, University of Waikato, Hamilton, New Zealand. 1994.

[Michie, 1991] Michie, D. Methodologies from machine learning in data analysis and software. *The Computer Journal,* 34(6): 559-565, 1991.

[Michie, 1991; Pazzani and Kibler 1992] Pazzani, M. and Kibler, D. The utility of knowledge in inductive learning. *Machine Learning*, 9(1): 57-94, 1992.

[Piatetsky-Shapiro and Frawley 1991] Piatetsky-Shapiro, G. and Frawley, W.J. *Knowledge discovery in databases*. AAAI Press, Menlo Park, CA, 1991.

[Quinlan, 1990] Quinlan, J.R. Learning logical definitions from relations. *Machine Learning,* 5: 239-266, 1990.

[Quinlan, 1991] Quinlan, J.R. Determinate Literals in Inductive Logic Programming. *Proceedings 12th International Joint Conference on Artificial Intelligence*, 746-750, 1991.

[Quinlan, 1993] Quinlan, J.R. and Cameron-Jones, R.M. FOIL: a midterm report. *Proceeding European Conference on Machine Learning,* p3-20, 1993.

[Witten et al., 1993] Witten, I.H., Cunningham, S.J., Holmes, G., McQueen, R., and Smith, L. Practical machine learning and its application to problems in agriculture. *Proceedings of the New Zealand Computer Society Conference,* pages 308-325, Auckland, New Zealand, 1993.