Classification and Visualization on the Hepatitis Dataset

Fabrizio Riguzzi

Dipartimento di Ingegneria, Università di Ferrara, Via Saragat 1
44100 Ferrara, Italy,
friguzzi@ing.unife.it

Abstract. In this paper we address goals 2 and 3 of those proposed by the donors of the Hepatitis dataset, namely to evaluate whether it is possible to estimate the stage of liver fibrosis from the results of examinations, and to evaluate the effectiveness of the interferon therapy. Goal 2 was addressed by learning various classifiers that predict the value of fibrosis from the values of examinations other than the biopsy. Unfortunately, the best accuracy obtained was only 50.6 %, up only 2.1 % from the performance of the default classifier, thus showing that replacing biopsies is still very hard if not impossible. As regards goal 3, we have plotted the distribution of the values of the difference in fibrosis and in activity before and after the interferon therapy. The plots show that the therapy actually reduces the level of activity but not the level of fibrosis. Moreover, we have also plotted the distribution of the values of the difference of GOT before and after the therapy. The graph shows that a moderate reduction of GOT is obtained.

1 Introduction

In this paper we present a knowledge discovery effort that addresses the goals 2 and 3 of those proposed by the donors of the Hepatitis dataset, namely:

2. evaluate whether laboratory examinations can be used to estimate the stage of liver fibrosis (if possible, we may be able to use laboratory examinations as the substitute for biopsy because biopsy is invasive to patients);
3. evaluate whether the interferon therapy is effective or not.

As regards the first goal, we first build a single table with a record for each biopsy that is preceded by at least one examination in the previous year. Each record stores the identifier of the patient, the date of the biopsy, the level of liver fibrosis detected in the biopsy and an attribute for each possible in-hospital examination.

From this table, a classifiers is learned that predicts the value of the fibrosis attribute on the basis of all the other attributes. A number of classifiers are learned: a decision tree, a majority classifier, a one rule classifier, a decision table, a conjunctive rule learner and a decision list learner. The learning is performed using the WEKA data mining system [5, 6].

As regards the second goal, we examine the value of a number of indicators on the patients on which the interferon therapy was applied. We consider the difference of the fibrosis before and after the interferon therapy (and analogously for activity). We graph the distribution of values for the two differences.

Moreover, we consider another indicator of the performance of the liver: GOT. This is an in-hospital examination. We consider the difference in GOT before and after the therapy. The distribution of the GOT difference values is also represented in a graph.

The elaboration of the data was performed using Microsoft Access and Microsoft SQL Server.

2 Estimation of Liver Fibrosis

The seven tables composing the dataset have been imported into an Access database. Then a number of elaborations have been done in order to obtain a table that reports, for each biopsy, the result of the biopsy regarding the liver fibrosis together with the results of all the in-hospital examinations done in the year preceding the biopsy. In the following we describe the queries that were run in order to obtain such a table.

First, we have extracted the patient identifier, the examination name and the examination date for all the examinations that were done in the year preceding the biopsy. Then, we have extracted, for each examination obtained from the previous query, the result of the examination (table `ExamsResBefBiopsy`).

The table resulting from the query has been materialized. Then it was manually cleaned. The table initially contained 29967 records, after cleaning it contained 29854 records. The cleaning involved two types of operations:

- the removal of examinations whose result contained a single word for examinations that are quantitative (e.g., D-BIL, TP),
- the removal of extra character from numerical results (e.g., P-QN had value '10>', '>' was removed; PSP15 had value '23.7 VL', 'VL' was removed).

In general, for a non-numerical result, if it was clear how to interpret the result in a numerical way, then the result was modified and kept. If there was no evident translation into a number, the result was discarded if the examination was quantitative and kept if the examination returns a nominal result.

We have then found the biopsies that have at least one examination in the year before together with their results and dates. This query results in 30028 records. Then duplicate records have been removed from the output table obtaining a table with 633 records.

A table `Exams` was created in Microsoft SQL Server that has fields `MID`, `Biopsy_date`, `Biopsy_fibrosis` and one field for each exam in `labn_e030704`. Thus the table has 459 + 3=462 attributes. Java (through JDBC) was used to create the table from the table `labn_e030704`. The type of the attributes is `FLOAT` for all except for those that in the cleaning phase or from the available

Algorithm	Accuracy
ZeroR	48.4992 %
OneR	46.7615 %
J48	50.5529 %
DecisionTable	48.4992 %
PART	49.6051 %
ConjunctiveRule	49.2891 %

Table 1. Mean accuracies over the 5 folds for the learning algorithms

information were identified as being nominal (a total of 33), that were declared as `VARCHAR(15)`.

A query was run in order to insert into the table `Exams` one record for each patient including the patient identifier, biopsy date, value of fibrosis and NULL values for all the examinations. Then Java (through JDBC) was used to update the tuples of `Exams` in order to replace the NULL values there with the results of examinations. For this purpose, for each record in `ExamsResBefBiopsy`, we have updated the table `Exams` by setting the field with the examination name equal to the examination result. Then the file was loaded in WEKA and the following classifier learning algorithms were applied: ZeroR, OneR, J48, DecisionTable, PART and ConjunctiveRule.

ZeroR returns the default classifier, i.e. the classifier that assigns to each instance the majority class in the training set. OneR returns a classifier with a single rule involving a single attribute, in other words, uses the minimum-error attribute for prediction, discretizing numeric attributes [2]. J48 is an implementation of Quinlan's c4.5 system [4]. DecisionTable is an algorithm that return a decision table. A decision table contains a number of labeled instances and it has a subset of the attributes that are present in the training set. A new instance is classified by looking in the table for an exact match over the attributes of the table. If no matching instances are found, the majority class is returned, if more than one instance is found, the majority class of all matching instances is returned [3]. PART returns a decision list. It uses separate-and-conquer. It builds a partial C4.5 decision tree in each iteration and turns the "best" leaf into a rule [1]. ConjunctiveRule returns a single rule with an antecedent consisting of a number of conditions conjoined together. If a test instance is not covered by the rule, then it is predicted using the majority class of the data not covered by the rule in the training data. This learner selects an antecedent by computing the Information Gain of each antecedent.

Table 1 reports the mean accuracies of the various learning algorithms computed using 5-fold cross validation. All the algorithm were run with the default parameters. As can be seen, the algorithm with the best accuracy is J48 but its accuracy is very close to the default accuracy of 48.4992 % (the one found by ZeroR).

Let us examine in more detail the decision tree that is found by J48 using all the instances as training set. The unpruned tree returned by j48 has 425 leaves and 476 nodes overall. It correctly classifies 342 instances out of 633 (accuracy on the training data of 54.0284 %). The pruned tree is the following

```
ALB <= 3.8: 4 (74.12/44.12)
ALB > 3.8: 1 (558.88/263.0)
```

for a total of 2 leaves and 3 nodes overall. It correctly classifies 326 instances for an accuracy of 51.5008 % on the training data, thus only slightly lower than the unpruned tree. Therefore we can conclude that the in-hospital examinations can not be reliably used to correctly predict the fibrosis of the liver.

3 Effectiveness of the Interferon Therapy

In order to investigate whether the interferon therapy was effective or not we have considered the results of biopsies before and after the end of the interferon therapy and the result of the GOT examinations also before and after the end of the interferon therapy.

We have extracted the difference in the levels of fibrosis and activity just before and just after the interferon therapy.

The resulting table contain 71 records. However, even if each patient was the subject of a single therapy, some patients appear more than once since they have more than one biopsy before the therapy start or after the therapy end. Therefore, we manually removed the extra records keeping only the biopsies that are closer to the therapy boundary dates. The resulting table has 37 records.

The table was then loaded into WEKA in order to count the occurrences of each value of the difference between the fibrosis levels before and after the therapy and of the difference between the activity levels before and after the therapy. Figure 1 shows the occurrence counts for the fibrosis difference while figure 2 shows the occurrence counts for the activity difference. From these graphs it can be observed that the interferon therapy has, on average, no effect on fibrosis but a certain effect on activity.

Moreover, we studied as well the differences of the GOT in-hospital examinations before and after the interferon therapy since GOT is an important indicator of liver health. We first obtained the dates of GOT examinations just before the start of the interferon therapy. We then extracted the results of GOT just before the start of the therapy. Afterwards, we obtained the dates of GOT just after the end of the therapy. Then we extracted the results of GOT just after the end of the therapy. Finally, we compute the difference between the two GOT examinations. Figure 3 shows the distribution of values for the difference of GOT. It has a minimum of -746, a maximum of 340, a mean of -32.492 and a standard deviation of 87.97. From the graph and the statistics we can conclude that the interferon therapy tends to reduce the value of GOT.

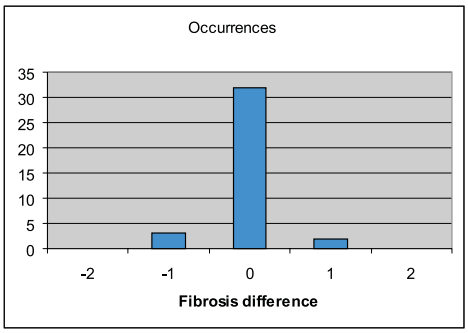

Fig. 1. Difference in fibrosis before and after the interferon therapy.

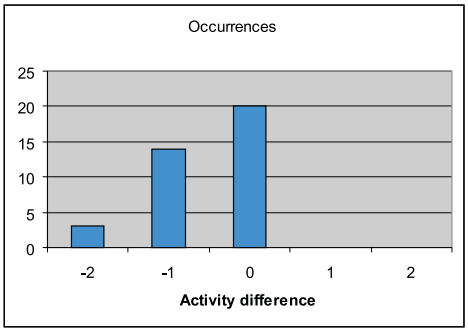

Fig. 2. Difference in activity before and after the interferon therapy.

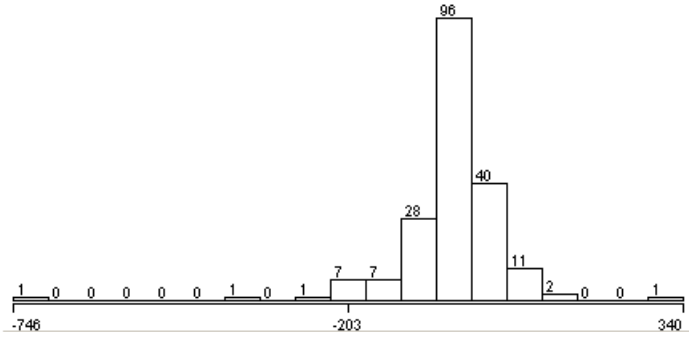

Fig. 3. Distribution of difference of GOT values before and after the interferon therapy.

4 Acknowledgements

This work is partially funded by the Ministero dell'Istruzione, della Ricerca e dell'Università under the COFIN2003 project "La gestione e la negoziazione automatica dei diritti sulle opere dell'ingegno digitali: aspetti giuridici e informatici".

References

1. Eibe Frank and Ian H. Witten. Generating accurate rule sets without global optimization. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers, 1998.
2. R. C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91, 1993.
3. R. Kohavi. The power of decision tables. In *Proc. European Conference on Machine Learning*, 1995.
4. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, USA, 1988.
5. I. H. Witten and E. Frank. *Data Mining - Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, USA, 2000.
6. I. H. Witten and E. Frank. WEKA, 2004. `http://www.cs.waikato.ac.nz/ml/weka/index.html`.