Machine learning methods to understand hepatocellular carcinoma pathology

Federica Ciocchetta
ITC-irst
Via Sommarive 18
38050 Povo (TN), Italy
E-mail: ciocchetta@itc.it

Francesca Demichelis
ITC-irst
Via Sommarive 18
38050 Povo (TN), Italy
E-mail: michelis@itc.it

Andrea Sboner
ITC-irst
Via Sommarive 18
38050 Povo (TN), Italy
E-mail: sboner@itc.it

Abstract

Hepatocellular carcinoma (HCC) is one of the commonest fatal tumors, and it is usually diagnosed at a late stage, when effective treatment is very difficult. Unfortunately early diagnosis of HCC is almost mandatory in terms of patient survival, but it represents a very difficult task. Detailed histological characteristics of small HCC and precursor lesions in histological diagnosis are needed to refine the contribution of histopathology to the management of patients with HCC. To enhance transparency and interpretability of the results, we applied non-blackbox machine learning algorithms on a set of 180 diagnosed cases of liver nodular lesions.

1 Introduction

Hepatocellular carcinoma (HCC) is one of the most prevalent tumors and causes of cancer death in the world, with about 1.25 million people died every year. This is due to the fact that HCC is often diagnosed at a late stage, when effective treatment is extremely critic and the prognosis poor. Therefore, the early diagnosis of small HCC is of great importance to guarantee effective treatment and efforts have to be made to accurately identify "early" or "small" lesions. Unfortunately, this is particularly hard, mainly because the specific histopathological and morphological criteria are uncertain and inadequate and the diagnosis of early lesions still depends on subjective interpretation. One of the most critical point is the histological differentiation between small HCC and precursor lesions.

In this context, machine learning methods can give an useful support to knowledge discovery.

Transparency and understandability of results play a great role in this kind of investigation by giving additional diagnostic models to clinicians. It is well known that, in the medical field, users require substantial explanation of model outputs, especially when face with unexpected solutions [7].

Only a few papers have been proposed previously for the diagnosis of HCC, such as *conjunctive normal form systems* [5] and *classification tree and neural network algorithms* [8].

The authors previously reported investigations regarding the classification of uncertain nodules, so called dysplastic nodules, and proposed a combination of five classifiers [3]. Besides, outputs of an unsupervised method was combined with the previous results, to support pathologists in the diagnosis of uncertain nodules [2].

2 Materials and Methods

2.1 Features and data

In this paper we considered 11 histological features, which are currently considered by the community of liver pathologists as the most useful histological criteria in the histological assessment of hepatocellular lesions, as it appears from published work on this subject (see for example [9],[4],[1]). However, in some dissertations on this topic, considerations have been done on the definition of a feature subset to enhance the standard diagnostic process [3].
The dataset was provided by a group of pathologists of Royal Free Hospital and University College Medical School of London. We have 180 lesions: 106 HCC (malignant lesions) and 74 macro-regenerative nodules (MRN - benign lesions). The nodules had been isolated during the routine diagnostic pathological examination of cirrhotic livers removed from 68 patients who received liver transplantation at the Royal Free Hospital between 1996 and 2001.

Table 1 reports the histo-pathological features: 7 features are categoric (4 nominal and 3 ordered) and 4 are numeric (size is expressed in millimeters). As usual in the medical domain, the diagnostic criteria play as malignancy indicator: higher feature values represent higher degree of malignancy. For completeness, we also report in table 1 the coding tables used by pathologists in lesion assessment. This coding table is currently used in hepatocellular carcinoma diagnostic process.

Table 1: List of features and their values

Features	Description	values
Nodule Size	real measurements - numeric	[mm]
Tumor necrosis	Absent/Present	0,1
Vascular Invasion	Absent/Present	0,1
Tumor Capsular Invasion	Absent/Present	0,1
Nodule Heterogeneity	Absent/Present	0,1
Reticulin Loss	Absent, mild, mild to moderate, moderate, moderate to severe, severe	0,1,2,3,4,5
Trabecular Thickness	number of liver cells forming trabecular	1,2,3,4,5,6,7
Capillarization	Marginal, patchy mainly marginal, patchy, incomplete, diffuse and incomplete, diffuse	0,1,2,3,4,5
Solitary Arterioles	number of arterioles per MPF	0,1,2,3,4,5,6,7
Cellular Atypia	Absent, mild, mild to moderate, moderate, moderate to severe, severe	0,1,2,3,4,5
Mitotic Activity	number of mitosis per HPF	0,1,2,3,4,5,6,7,8,9,10,11,12

2.2 Methods

From the clinical point of view very strong user requirements are the interpretability of the model and prediction accuracy. Moreover, the definition of the most relevant features could be of great interest.

In our case, the requirements of transparency and understandability of the model itself drove the choice of the learning algorithms: a model-based learning approach by using a rule induction algorithm and a decision tree inducer were implemented. Furthermore, the taken approach could address a feature selection task, in some sense, by suggesting pathologists the more informative diagnostic criteria among the whole feature set. This aspect could provide insights into the underlying biological process that leads to HCC.

In our investigations we adopted 10-fold cross validation, to avoid model overfitting on the training data. As regards performances evaluation, in addition to accuracy, we used sensitivity and specificity, as commonly done in medical field.

Rule induction algorithm The task of rule induction is: **Given** a set of classified *examples*, **find** a set of classification *rules* that are *accurate* and *informative*.
In particular, we used the CN2 algorithm implemented in Weka package [6] by Institute Jožef Stefan in Ljubljana. CN2 uses the covering approach to construct a set of rule for each possible class c_i in turn [7]. An IF-THEN rule, in classification tasks, is defined as: IF *Conditions* THEN *Prediction*.
The models created by this method have the great advantage to be immediately interpretable by the user.

Decision tree algorithm We applied C4.5 decision tree learner implemented in Weka package as J48 classifier. One of the problems involved with decision tree classifier is finding the right dimension of the tree. In fact, a too short decision tree could result in poor prediction, while a too long decision tree could be an optimal tree for the training set, but a worse one for new cases. To solve this problem, generally, the tree is pruned after being fully expanded (post pruning or backward pruning).
In addition it is worth pointing out that a decision tree can be seen as a collection of rules, with each terminal node corresponding to a specific decision rule.

3 Results and evaluation

3.1 Rule induction

Regarding the rule induction algorithm, we employed the Laplace estimate and its generalization, the m-estimate, in rule evaluation. We performed the m-estimate with different values of m. The parameter m controls the role of the prior probabilities and the evidence provided by the examples: higher m gives more weight to the prior probabilities and less to the examples. Therefore, higher values of m are appropriate for noisy dataset. To refer to different experiments we coded each of them with a letter, indicating the type of estimate (a, b, c, d respectively for Laplace, m=2,m=1 and m=0) and an index indicating the value of *maxStarSize* (see table 2), which, in some sense, tunes the rule complexity. For example, regarding the rules induced by the algorithm with *maxStarSize*=5, we obtained two sets of rules for models *a.0* and *b.0* and for models *c.0* and *d.0*, respectively:

i. IF Reticulin-loss = 0 AND Heterogenity = 0
THEN Diagnosis = MRN [0,68]
ELSE IF Trabecular-Thick > 3.5 THEN Diagnosis = HCC [62,0]
ELSE Diagnosis = HCC [44,6]

Table 2: Rule induction performances - complexity

Model	*max Star Size*	**Rule evaluation**	**N. of rules**	**Accuracy** [%]	**Sensitivity** [%]	**Specificity** [%]
a.0	5	Laplace	3	96.1	98	93
b.0	5	m-estimate (m=2)	3	96.1	98	93
c.0	5	m-estimate (m=1)	4	96.1	98	93
d.0	5	m-estimate (m=0)	4	96.7	97	96
a.1	10	Laplace	4	95.6	95	96
b.1	10	m-estimate (m=2)	4	95.6	95	96
c.1	10	m-estimate (m=1)	4	95.6	95	96
d.1	10	m-estimate (m=0)	4	95.6	95	96
a.2	0	Laplace	3	96.1	95	97
b.2	0	m-estimate (m=2)	3	96.1	95	97
c.2	0	m-estimate (m=1)	3	95.6	94	97
d.2	0	m-estimate (m=0)	8	98.3	97	100

ii. IF Capilarization = 0 AND Reticulin-loss = 0 THEN Diagnosis = MRN [0,68]
ELSE IF Trabecular-Thick > 3.5 THEN Diagnosis = HCC [62,0]
ELSE IF Solitary-Arterioles < 0.5 AND Atypia = 0 THEN Diagnosis = MRN [0,5]
ELSE Diagnosis = HCC [44,1]

Note that even with different set of rules, accuracy, sensitivity and specificity do not present a great variation across the models (models *a*, *b* and *c* have identical values). Following the Occam's razor principle, models with simpler rules (i) should be preferred. We performed the same experiments with different values of *maxStarSize* (equal to 0 and 10).

3.2 Decision tree

We applied C4.5 algorithm to induce the tree. We induced binary trees, applying different options to prune them. The first experiment built the unpruned tree, the second used a confidence factor with the default value of 0.25 as pruning option, and the third one used a holdout set to perform pruning. The latter uses a standard verification technique to estimate the error. However, it suffers from the disadvantage that less data are used to build the tree. On the contrary, the confidence factor is a way to estimate the error directly from the training data. The less the confidence factor, the deeper the pruning. Table 3 and figure 1 report the results.

We then modified the options for the decision tree trying to improve the performances. Among the different available options, we built several classifiers with various confidence factors for pruning. For values greater than 0.50 we obtained the unpruned tree (see fig. 1(a)), while for confidence factor ranging from 0.09 to 0.50 there is no changes with respect to the results obtained with the default value 0.25. Finally, we obtained a shorter tree but also poorer performances with confidence factors less than 0.09 (accuracy = 95.0; sensitivity = 95; specificity = 95). Analogously, by using the holdout method to prune the tree, we modified the number of the folds (from 2 to 10). The procedure yielded the same results as given by the model *g* with default value equal to three (see table 3 and fig. 1(c)).

3.3 Comparison

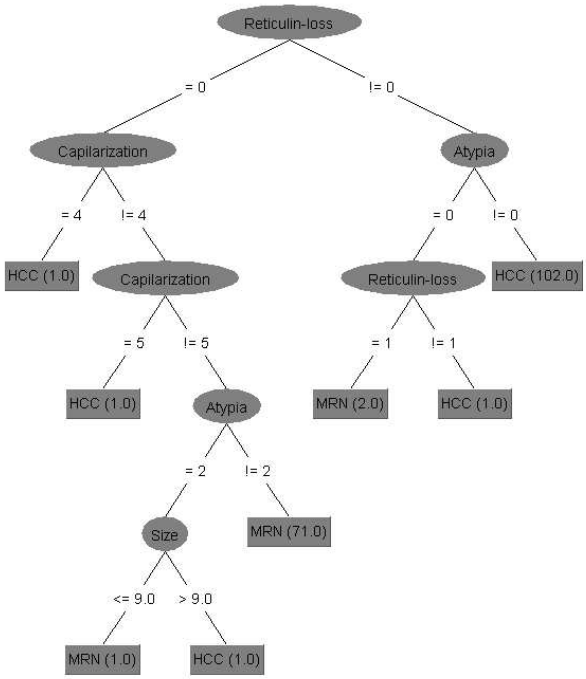
We compared classifiers performances by means of a paired t-test, which showed no statistical differences among them with some exceptions (p-value < 0.05 for significance). Exceptions are mostly due to the decision tree with the hold-out method. Anyhow, there is no classifier that outperforms all others in terms of accuracy, sensitivity, and specificity. Therefore we cannot state which classifier is more suitable to our purpose. In terms of the feature subsets selected by the classifiers, we note that:

- Reticulin Loss has been always selected (12 rules and 3 trees).
- Trabecular Thickness, selected in 12 rules;
- Cellular Atypia, selected in 7 rules and 2 trees;
- Capillarization, selected in 6 rules and 2 trees;
- Solitary Arterioles, Nodule Heterogeneity, selected in 6 rules;
- Size, selected in 1 rule and 1 tree;

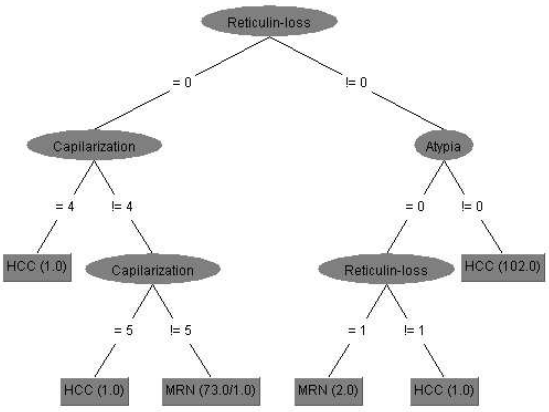
Four features (Tumor necrosis, Vascular Invasion, Tumor Capsular Invasion, and Mitotic Activity) have never been included. Note that Size has been selected only for the unpruned tree and the rule with $maxStarSize = 0$, while Reticulin Loss is always selected as the root node of the trees.

4 Discussion

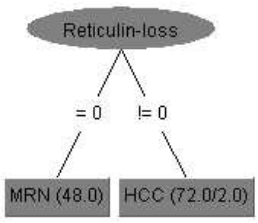
No validate histological diagnostic criteria are now available to diagnose early hepatocellular carcinoma. Few papers proposed the use of computerised systems



(a) Unpruned tree



(b) Standard pruning



(c) Pruning with holdout set

Figure 1: Decision trees

to help pathologists. Di Giacomo et al. [5] approached this task by using logic reasoning in terms of *conjunctive normal form systems*; they did not provide any validation of their outcomes. Poon et al. [8] compared *classification tree* and *neural networks* for the identification of serological liver marker profiles, reporting similar diagnostic values in differentiating the two classes. To our knowledge no scientific paper faces the classification problem from a histo-pathological viewpoint in terms of computerised systems.

In this report, machine learning methods were applied to support the early diagnosis of hepatocellular carcinoma and to get relevant insights into this disease. It is well know that the development of a computerised decision support system in the medical field should take into account the interpretability of the

Table 3: Decision tree

Model	Description	Tree Size	N. of leaves	Accuracy [%]	Sensitivity [%]	Specificity [%]
e	Unpruned	15	8	96.1	95	97
f	CF 0.25	11	6	96.1	95	97
g	Holdout	3	2	97.2	97	97

model, besides its accuracy [7].

Due to these requirements, among different machine learning algorithms for classification, we decided to apply two symbolic methods, namely a rule induction algorithm (CN2) and a decision tree learner (C4.5 - J48). We evaluated their performances in terms of accuracy, sensitivity and specificity by applying a stratified 10-fold cross validation. For each method, we performed different experiments tuning the parameters of the models in order to find the optimal trade-off between prediction accuracy and interpretability of the results.

As a general comment, both algorithms perform an intrinsic "feature selection" procedure in some sense. As a matter of fact, every experiment we conducted, used only a subset of the whole feature set to create the model.

Regarding the rule induction algorithms, slightly better performances have been reached with model *d.2* (see table 2). This model implements *maxStarSize* parameter equal to 0 and accuracy as rule evaluation method ($m = 0$). In this case, only **one** attribute is used in the IF part of the rule and a total of five attributes makes up the classifier. In some sense, interpretability is penalized with respect to other models, due to the greater number of rules.

A further consideration regards the m value for the m-estimate. No relevant variations have been detected in our experiments; we guess that it can be addressed to the relatively not noisy dataset.

Similar results in terms of accuracy, sensitivity, and specificity have been obtained with decision tree classifiers. It is worth noting that all the attributes selected by the trees were also used by the CN2 classifiers. Moreover, for the pruned trees, the selected attributes are comparable to those identified in a previous work on HCC data[3]. In that paper, feature selection algorithms, employing filtering methods (Correlation Based Feature Selection and Relief), were applied to reduce the dimensionality of the feature space by selecting the more relevant attributes, to enhance subsequent classification task.

As a general result, four features have never been selected, suggesting their "irrelevance" for diagnostic purposes. On the contrary, Reticulin Loss is giving as the most discriminant feature.

We remark that extracting more knowledge about the nature of hepatocellular carcinoma nodules is clinically relevant in term of patient's prognosis and treatment. Currently a set of eleven features is considered by many liver pathologists as the most useful histological criteria. However, no definitive agreement has been found yet in pathologist community. In this context, it is our opinion that the above reported results could help pathologists in their diagnostic decision-making process and, at the same time, could provide new knowledge on the disease.

Unfortunately, no comprehensive studies focused on the analysis of pathological data from a machine learning approach are yet reported in literature.

As regards future work on this topic, as a first step a greater number of data are needed to perform more reliable data analysis. Moreover, we would like to point out that in Ciocchetta et al.[3] attention has been given to the so-called *dysplastic nodules*: nodules whose diagnosis is uncertain. Therefore, owing to those considerations, the methods we applied in this report will be used to classify dysplastic nodules.

Another step to be done regards the involvement of pathologists in nodule evaluation: thanks to the transparency of the models we built, we asked them to re-evaluate those uncertain lesions, following our results. To really evaluate the validity of a model a continuous feedback from pathologists and clinicians has to be promoted.

Acknowledgments

The authors thank Amar Paul Dhillon and Alberto Quaglia of the Royal Free Hospital and University College Medical School.

References

[1] Anonymous. Terminology of nodular hepatocellular lesions. international working party. *Hepatology*, 22:983–993, 1995.

[2] F. Ciocchetta, R. Dell'Anna, F. Demichelis, A. P. Dhillon, A. Quaglia, and A. Sboner. Combining supervised and unsupervised methods to support early diagnosis of hepatocellular carcinoma. In *accepted for oral presentation at AIME 2003*, 2003.

[3] F. Ciocchetta, R. Dell'Anna, F. Demichelis, A. Sboner, A. P. Dhillon, A. Dhillon, A. Godfrey, and A. Quaglia. Knowledge discovery to support hepatocellular carcinoma early diagnosis. In *International Joint Conference on Neural Network - Special Session: Knowledge Discovery, and Image and Signal Processing in Medicine*, 2003.

[4] L. D. Ferrell, J. M. Crawford, A. P. Dhillon, P. J. Scheuer, and Y. Nakanuma. Proposal for standardized criteria for the diagnosis of benign, borderline, and malignant hepatocellular lesions arising in chronic advanced liver disease. *Am J Surg Pathol*, 17(11):1113–23, 1993.

[5] P. D. Giacomo, G. Felici, R. Maceratini, and K. Truemper. Application of a new logic domain method for the diagnosis of hepatocellular carcinoma. In *Proceedings of MEDINFO 2001*. Amsterdam: IOS Press, 2001.

[6] I. H.Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[7] N. Lavrač. Selected techniques for data mining in medicine. *Artif Intell Med*, 16(1):3–23, 1999.

[8] T. C. W. Poon, A. T.-C. Chan, B. Zee, S. K.-W. Ho, T. S.-K. Mok, T. W.-T. Leung, and P. J. Johnson. Application of classification tree and neural network algorithms to the identification of serological liver marker profiles for the diagnosis of hepatocellular carcinoma. *Oncology*, 61:275–283, 2001.

[9] A. Quaglia, S. Bhattacharjya, and A. P. Dhillon. Limitations of the histopathogical diagnosis and prognostic assessment of hepatocellular carcinoma. *Histopathology*, 38:167–174, 2001.