CLASSIFICATION: CERVICAL CANCER(BIOPSY) ON THE BASIS OF THE RISK FACTORS

Harpragaas Singh
(U6424131)
The Australian National University

ABSTRACT

This study presents a data set that comprises demographic information, habits medical records of 858 information.The dataset focuses on the prediction of indicators of cervical cancer. In this report, we have only considered one of the indicators, Biopsy.[1]
We look at a method for generating a high level of classification that can be used to generate a more accurate result and compare it with a paper that cites this data set. The classification technique used is the Decision Tree Classification technique.
We describe a simple method to implement a neural network. We train and test the dataset in the neural network to calculate and compare the accuracy of the predictions.[2,3]

INTRODUCTION

Despite the possibility of prevention with regular cytological screening, cervical cancer remains a significant cause of mortality.This being the cause of more than half a million cases per year and killing more than a quarter of a million in the same period [1,4]. As in many other diseases, the existence of several screening and diagnosis methods creates a complex ecosystem from a Computed Aided Diagnosis (CAD) system point of view. For instance, in the detection of precancerous cervical lesions, screening strategies include cytology, colposcopy (covering its several modalities [1,4]) and the gold-standard biopsy. In developing countries, resources are very limited and patients usually have poor adherence to routine screening due to low problem awareness. Consequently, the prediction of the individual patient's risk and the best screening strategy during her diagnosis becomes a fundamental problem

In this paper, we will generally assume a feed-forward network of three layers of processing units namely input, output, and hidden layers. All connections are from units in one level to the subsequent one, with no lateral, backward or multilayer connections.While defining the neural network the number of input neurons is equal to the number of input attributes of the dataset,(which is seven in this case), and the number of output neurons equals the number of classes the output data has divided into(which is in boolean in this dataset) and only one target attribute 'Biopsy' has been considered.

THE DATA

The Data was collected at Hospital Universitario de Caracas' in Caracas, Venezuela.The datasets include records from 858 patients who were asked questions on their respective sexual lives. However, Several patients decided not to answer some of the questions because of privacy concern .[1]

DATA PREPROCESSING

The data has been preprocessed by using the Sensitive Analysis technique[5,6]. This technique helps us to understand the effect an input neuron has on an output neuron by determining the effect this change will have on the output. If the change is drastic and the input neuron is considered to be a key factor in producing current activation function of the output neuron. Input patterns are identified by observing its effect on output.

To produce concise, understandable explanations the following methodology has been followed[5]

1. Liken the input pattern to the characteristic input patterns, and present the most similar to the user.
2. In addition present inputs considered 'important' for the current network output, and their values in the characteristic pattern.
3. The hidden values have been identified and removed.
4. Produce a set of rules, and evaluate to confirm accuracy.
5. Give the network's next most likely output.

DECISION TREE CLASSIFICATION-ID3 ALGORITHM

To produce the set of rules for classifying on the basis of the indicator we use the C4.5 algorithm. This program includes the ID3 algorithm which is a machine learning algorithm for building classification trees developed by Ross Quinlan in/around 1986. The algorithm is a greedy, recursive algorithm that partitions a data set on the attribute that maximizes information gain.[3]

To produce a decision tree, the class attribute(target attribute) is considered to be the **leaf node**. The entropy of the target attribute is calculated using,

$$H(S) = \sum_{x \in X} -p(x) \log_2 p(x) \qquad (1)$$

Here, 1. S is the class attribute
2. H(S) is Entropy of the class attribute

Then for each attribute, information gain is calculated using,

Information Gain= Entropy of Class Attribute- Entropy of (each) attribute

$$IG(A, S) = H(S) - \sum_{t \in T} p(t) H(t) \qquad (2)$$

To construct the decision tree we calculate the entropy for the class attribute and entropies for all other attributes. Then we calculate the information gain for each attribute. The one with the highest gain is considered as the **root node.**

Example of the Decision Tree Classification

```
IF Num of pregnancies EQUALS 4 AND Age EQUALS 0 AND Dx EQUALS 0 AND Hormonal Contraceptives EQUALS 0 AND STDs EQUALS 1 AND IUD EQUALS 0 AND Smokes EQUALS 1 THEN 0
IF Num of pregnancies EQUALS 11 THEN 0
IF Num of pregnancies EQUALS 3 AND Dx EQUALS 1 AND IUD EQUALS 1 THEN 1
IF Num of pregnancies EQUALS 3 AND Dx EQUALS 1 AND IUD EQUALS 0 THEN 0
IF Num of pregnancies EQUALS 7 THEN 0
IF Num of pregnancies EQUALS 4 AND Age EQUALS 1 AND IUD EQUALS 0 AND Hormonal Contraceptives EQUALS 1 AND Smokes EQUALS 1 AND STDs EQUALS 1 THEN 1
IF Num of pregnancies EQUALS 1 AND STDs EQUALS 0 AND Dx EQUALS 0 AND Smokes EQUALS 0 AND Age EQUALS 0 AND Hormonal Contraceptives EQUALS 1 AND IUD EQUALS 1 THEN 0
IF Num of pregnancies EQUALS 8 THEN 0
IF Num of pregnancies EQUALS 2 AND STDs EQUALS 0 AND IUD EQUALS 0 AND Hormonal Contraceptives EQUALS 1 AND Age EQUALS 0 AND Dx EQUALS 1 THEN 0
IF Num of pregnancies EQUALS 4 AND Age EQUALS 0 AND Dx EQUALS 0 AND Hormonal Contraceptives EQUALS 1 AND Smokes EQUALS 1 AND IUD EQUALS 0 AND STDs EQUALS 0 THEN 0
```

Fig1. A sample of Decision Tree generated using id3.py

Comparison of the Accuracies

The following conditions are implemented to train and test the dataset to predict the accuracy.
To check and compare the accuracy

1. In this case, first 250 patterns are trained while 12 patterns out of 600-612 are trained which gave an accuracy of 99.2 % and 95% respectively.

2. In this case, the range of the training set was reduced to first 250 patterns and the range of the testing set is changed to 400 to 600 patterns which gave an accuracy of 98.99 % and 96.67% respectively

3. In this case, the rage of the training and testing set to first 300 patterns, this gave an accuracy of 98.66% each.

It is relatively unusual to achieve a higher value on the test set. This may be due to a higher proportion of noisy patterns in the training set.[2]

EVALUATION

LOSS FUNCTION

The accuracy of the neural network can be evaluated by comparing the loss function.
The lower the loss**,** the better a model (unless the model has over-fitted to the training data). The loss is calculated by training and validation and its interpretation is how well the model is doing for these two sets. Unlike accuracy, loss is not a percentage. It is a summation of the errors made for each example in training or validation sets.Here the loss function implies how well or poorly a certain model behaves after each iteration of optimization. Ideally, one would expect the reduction of loss after each, or several, iteration(s).

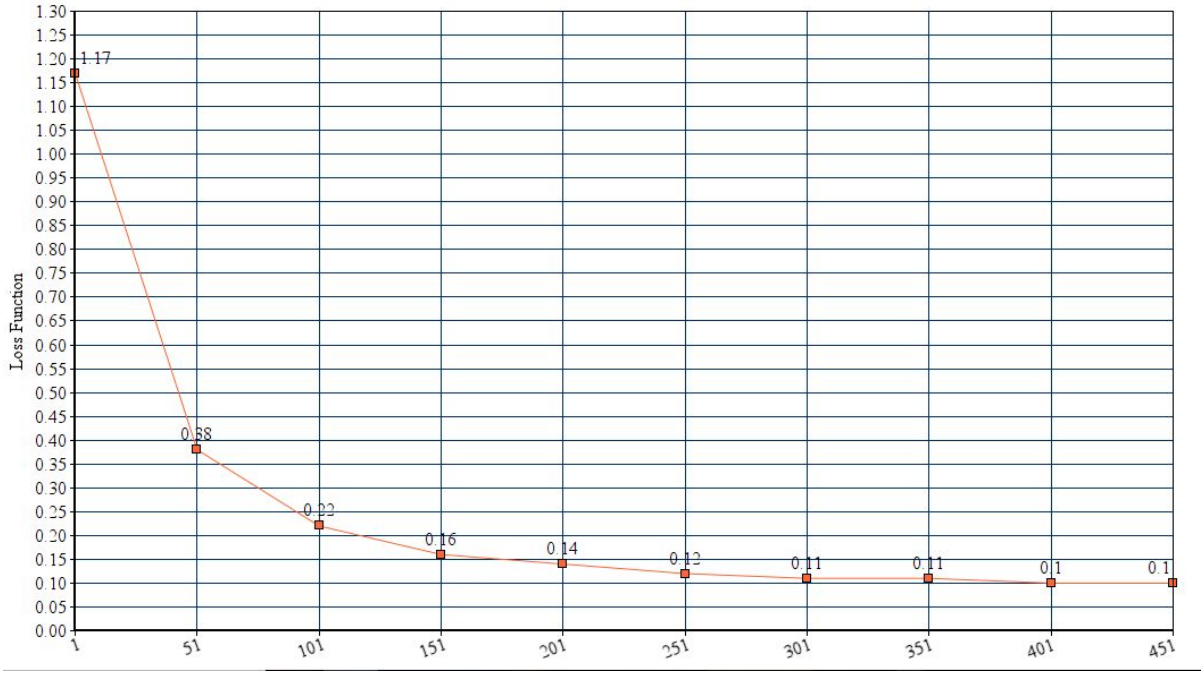

Fig2. Graph of the Loss Function

CONFUSION MATRIX

To see how well the network performs on different categories, we will create a confusion matrix.
The confusion matrix shows the ways in which your classification model is confused when it makes predictions.
In a confusion matrix, the row represents the predicted data while the column represents the actual data. The principal diagonal of the matrix shows the correct matches while training and testing the dataset

```
Confusion matrix for training:

 295    0    0
   4    0    0
   0    0    0
[torch.FloatTensor of size 3x3]

task1_hp.py:356: UserWarning: Implicit dimension choice for softmax has been de
precated. Change the call to include dim=X as an argument.
  _, predicted_test = torch.max(F.softmax(Y_pred_test), 1)
Testing Accuracy: 90.00 %

Confusion matrix for testing:

 9  0  0
 1  0  0
 0  0  0
[torch.FloatTensor of size 3x3]
```

Fiq 3. The Confusion Matrix where range is [1 to 300] for training the dataset and [590:600] for the testing dataset

In Fig 3 , we can see that in the principal diagonal of the confusion matrix for training set out of first 300 patterns 295 matches, in the principal diagonal of the confusion matrix for the testing set out of the 10 patterns 9 matches,

CONCLUSION AND FUTURE WORK

In this paper, the data can be classified to predict the risk factors that may cause cervical cancer.For simplicity ,the target attribute *here is only Biopsy thus the accuracy when we consider the other three target variable may vary.*
However, ID3 does not guarantee an accurate solution as it may get stuck in the local optima. It is a greedy, recursive algorithm by selecting the best attribute to split dataset on each iteration.Smaller decision trees are preferred over larger ones. This algorithm usually produces small trees, but it does not always produce the smallest possible tree.Thus, the predicted accuracy can be verified and improved by using other classification techniques.[4]

REFERENCES

1. Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.c
2. Bustos, R. A., & Gedeon, T. D. (1995). Decrypting Neural Network Data: A GIS Case Study. In Artificial Neural Nets and Genetic Algorithms (pp. 231-234). Springer, Vienna.
3. Quinlan, JR C4.5: Programs for Machine Learning, Morgan & Kaufmann, 1993.
4. Fernandes, K., Cardoso, J.S., Fernandes, J.: Temporal segmentation of digital colposcopies. In: Pattern Recognition and Image Analysis. Springer (2015) 262–271
5. Gedeon, T. D., & Turner, S. (1993, October). Explaining student grades predicted by a neural network. In Neural Networks, 1993. IJCNN '93-Nagoya. Proceedings of 1993 International Joint Conference on (Vol. 1, pp. 609-612). IEEE
6. Klimasauskas, CC "Neural networks tell why," DD Jour., April 1991.
7. ID3 ALGORITHM program URL: https://github.com/tofti/python-id3-trees