Applications of Decision Tree in Protein Relative Solvent Accessibility Prediction

Vineet Nayak[1, *, +], **Rahul Alapati**[1, +], **Anuj Gupta**[1, +], **Sritika Chakladar**[1, +] **and Austin Ream**[2, +]

[1]Graduate Student, Computer Science and Software Engineering, Auburn University, 36830, USA [2] Undergraduate Student, Computer Science and Software Engineering, Auburn University, 36830, USA

[*] vvn0001@auburn.edu

[+] these authors contributed equally to this work

ABSTRACT

Protein relative solvent accessibility (RSA) provides insight into understanding the protein structure and function. Prediction of protein relative solvent accessibility is often the first stage of predicting other protein properties. Decision Trees are quite helpful in accurately predicting the relative solvent accessibility of amino acid residues. This report lays out the implementation details of a popular decision tree learning algorithm namely: ID3 (Iterative Dichotomiser 3) algorithm. The ID3 algorithm is used to generate a decision tree from a database consisting of protein sequences and their respective relative solvent accessibilities. Precision, Recall, Accuracy, F-1 Measure and Mathews Correlation Coefficient (MCC) are calculated to determine the accuracy of the decision tree classifier generated using the ID3 algorithm.

1. Introduction

Solvent accessibility can be defined as, the size of the area of an amino acid that is exposed to the solvent (water). Maximum solvent accessible area for each amino acid is its whole surface area. The chemical properties of amino acid residues play an important role in determining the solvent accessibility. Hydrophobic residues like to be buried inside (interior). Hydrophilic residues like to be exposed on the surface.

Solvent accessibility is considered as an important measure of spatial arrangement during the process of protein folding. The solvent accessibility of an amino acid in a protein defines its surrounding solvent environment and hydration properties, this characteristic has been widely used to analyze protein structure and function. Prediction of relative solvent accessibility is often the first predictive stage of determining protein structure and function. Predicted RSA assists in

predicting protein secondary structure, domain boundary, disorder and hot spot, as well as protein-protein interaction prediction and fold recognition.

Knowledge of a protein's 3D structure can help us calculate the solvent accessible surface area for each amino acid. We can then calculate the relative solvent accessibility (RSA) by calculating what percentage of the amino acid's surface is exposed to the solvent. If the RSA is less than 25% then solvent accessibility can be determined as buried or else exposed.

Decision Trees can be used to efficiently predict the RSA of amino acid residues, based on their chemical properties. Decision trees are supervised algorithms which recursively partition the data based on its attributes, until some stopping condition is reached. This recursive partitioning gives rise to a tree-like structure. Decision trees are white boxes as the classification rules learned by them can be easily obtained by tracing the path from the root node to each leaf node in the tree.

Decision trees are very efficient even with the large volumes of data. This is due to the partitioning nature of the algorithm, each time working on smaller and smaller pieces of the dataset and the fact that they usually only work with simple attribute-value data which is easy to manipulate. The decision tree classifier is one of the possible approaches to multistage decision making. The most important feature of Decision Tree Classifier's is their capability to break down a complex decision–making process into a collection of simpler decisions, thus providing a solution which is often easier to interpret.

ID3 algorithm is invented by Ross Quinlan and it is the precursor to the C4.5 algorithm, and it is typically used in the machine learning and natural language processing domains. In decision tree learning, ID3 algorithm is used to generate a decision tree from a dataset. The ID3 decision tree algorithm follows a greedy top-down growth of the tree using information gain to learn the best hypothesis on training dataset.

2. Methods

The protein sequences in Multi-FASTA format, the true binary relative solvent accessibility labels (‘e’, ‘-‘) of the proteins and the chemical properties of the amino acid residues have been used to train and test the decision tree, built using the ID3 Algorithm.

The input protein sequences and their respective RSAs have been curated into non-overlapping sets Training (75%) and Test (25%) datasets using simple random sampling without replacement. The random sampling without replacement has been implemented using the random.seed() & random.random() methods of the random python library.

The random.seed() method is used to initialize a pseudorandom number generator and the random.random() method is used to return the next random floating-point number. We use this random number to curate the training (75%) and test (25%) datasets.

Now, a feature vector X has been constructed for the training and test datasets using the chemical properties of the 20 naturally occurring amino acids. We ignored the amino acid X and its respective RSA in both the training and test datasets.

The feature set consists of the following binary attributes:
Hydrophobic, Polar, Small, Proline, Tiny, Aliphatic, Positive, Negative, Charged.
The binary output labels Y are ‘e’ (1) and ‘-‘ (0).

Now, the problem setting is that of function approximation. We have a set of possible instances X i.e. the feature vector consisting of the chemical properties of amino acids. We also have a set of binary valued output labels Y. We need to develop a hypothesis h belonging to H, that best approximates the unknown function f, which uses X and produces Y.

Here, we have used the ID3 decision tree learning algorithm to solve the function approximation problem.

2.1. Decision Tree Learning using ID3 on Training Set:

The ID3 algorithm begins with the original set S as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set S and calculates entropy or information gain of that attribute. It then selects the attribute which has the smallest entropy or largest information gain value. The set S is then split by the selected attribute to produce subsets of data. The algorithm continues to recurse on each subset, considering only attributes never selected before.

Recursion on a subset may stop in one of these cases:

- Every element in the subset belongs to the same class (+ or -), then the node is turned into a leaf and labelled with the class of the examples
- There are no more attributes to be selected, but the examples still do not belong to the same class (some are + and some are -), then the node is turned into a leaf and labelled with the most common class of the examples in the subset.
- Examples still do not belong to the same class (some are + and some are -) but none of the remain attributes can split the dataset further i.e. max information gain is 0.
- There are no examples in the subset, this happens when no example in the parent set was found to be matching a specific value of the selected attributes. Then a leaf is created and labelled with the most common class of the examples in the parent set.

Throughout the algorithm, the decision tree is constructed with each non-terminal node representing the selected attribute on which the data was split, and terminal nodes representing the class label of the final subset of this branch.

After building the decision tree, we observed that the training dataset is not perfectly classified into either exposed or buried. Thus, we came up with a probability measure to determine the class i.e. exposed or buried for every path in the decision tree. The probability measures the chances of an amino acid being exposed if a particular path is followed in the resulting decision tree.

We then performed an analysis to determine the threshold value for the probability. We started from 0.0 and went up to 1.0 in steps of 0.1 and analyzed the F-1 Measure and MCC (as shown in Figure 1). The probability threshold of 0.4 resulted in maximum F-1 Measure and MCC values. Hence, after rigorous analysis 0.4 was selected as the probability threshold.

In our decision tree, if the probability threshold is greater than or equal 0.4, then the Amino Acid residue is considered to be “exposed”, otherwise “buried”.

2.2. Decision Tree Classification on Test Set:

Here, we have used the test dataset to walk over the trained tree generated using the ID3 algorithm. We, also predicted the output labels for each amino acid residue on the test dataset.

Using the predicted output labels and the original output labels in the test data, we calculated the number of True Positives, True Negatives, False Positives and False Negatives.

We used the following table to determine True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

		True Condition (Output Label from Test Data)	
Predicted		Exposed	Buried
output	Exposed	TP	FP
label	Buried	FN	TN

Now, we used the number of TP, TN, FP & FN to evaluate the accuracies as follows:

Precision = (TP)/ (TP + FP)

Recall = (TP)/ (TP + FN)

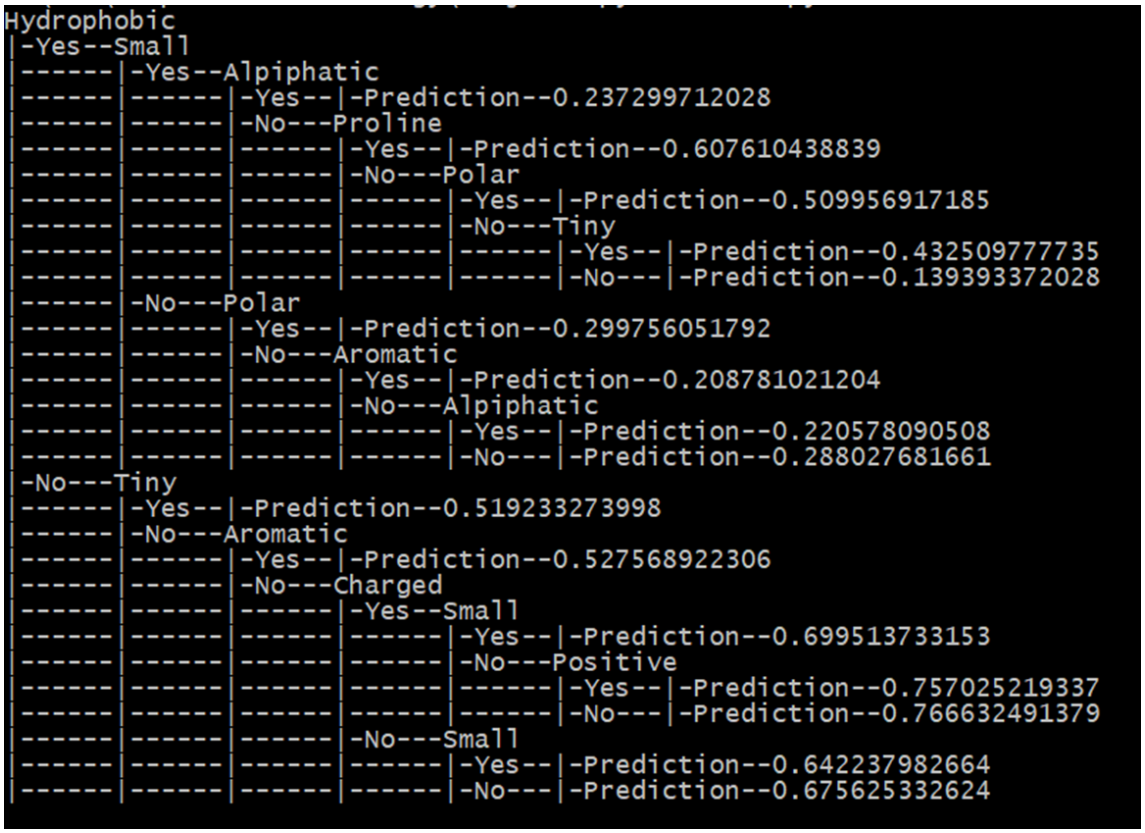
Accuracy = (TP + TN)/ (TP + TN + FP + FN)

F-1 Measure = 2 * ((Precision *Recall) / (Precision + Recall))

MCC = ((TP * TN) – (FP * FN)) / math.sqrt((TP+FP) (TP+FN) (TN+FP) (TN+FN)

3. Results

The following is our decision tree built on the training dataset using the ID3 Algorithm:



The prediction values in the decision tree are the probabilities of an amino acid being exposed, if it takes that particular path in the decision tree. For example, if an amino acid is Hydrophobic, Small and Aliphatic, then the probability that it is exposed is 0.237. We consider this amino acid to be buried as we have set our probability threshold to 0.4 as mentioned above.

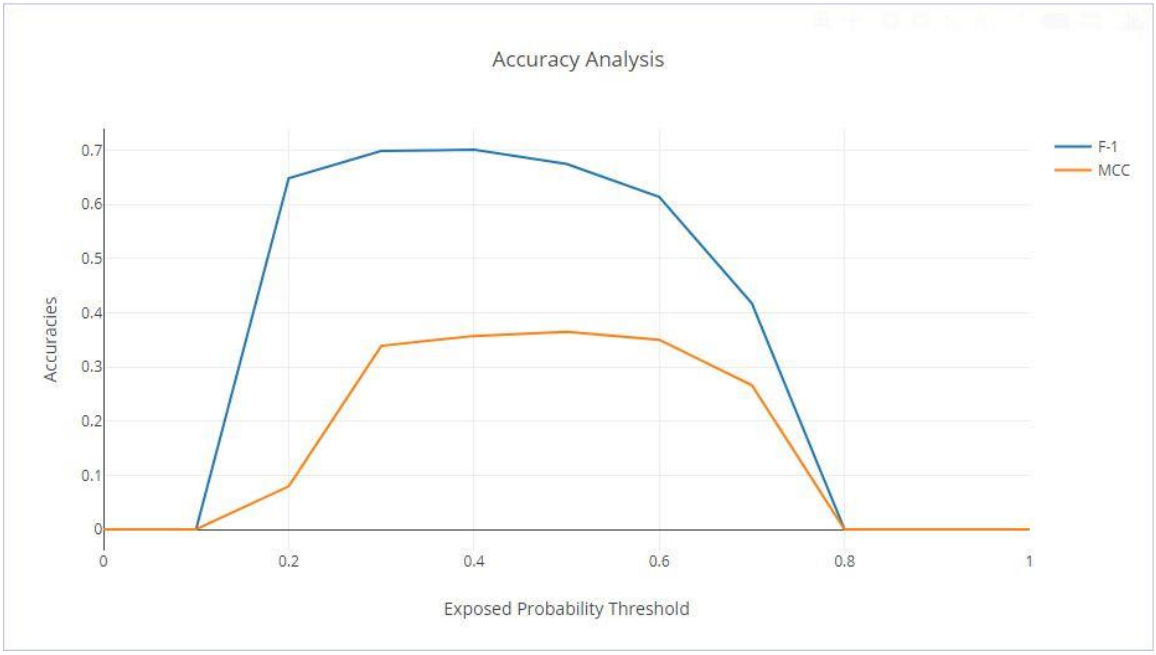


Figure 1: Exposed Probability Threshold Analysis

The following are the accuracy values of the decision tree classifier:

```
========= Accuracy ===========
Precision:0.602837508575
Recall:0.834793789105
Accuracy:0.661377753996
F-1 Measure:0.700103021978
Matthew's Correlation Coefficient (MCC):0.357514904881
```

Our decision tree has an accuracy of 66.13 %, F-1 Measure of 70.00 % & our predicted labels have a moderate correlation of 0.357 with the true labels. The results are significantly constant over multiple runs and curations of the input raw data.

4. Discussion

We tried different possibilities for determining the class (i.e. exposed or buried) as the data was not perfectly classified. We tried to use majority as a factor to determine the solvent accessibility, but it resulted in lower accuracy, F-1 Measure and MCC. We observed that in some cases, majority is biased. For example, if we take a particular path in the decision tree and it results in 6 cases out of 10 where the amino acid residue is exposed and in the other 4 cases it is buried. In cases, like these majority can be biased. Hence, we came up with the concept of probability and benchmarked its threshold. The probability measure resulted in greater accuracy, F-1 Measure and MCC.

Also, on many instances we could observe different solvent accessibilities for the same amino acid in a protein, although the attributes are the same. The accuracies of the decision tree classifier can be increased further if we optimize our training data, by removing the noise.

5. References

1. Nguyen MN1, Rajapakse JC. Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins. 2005 Apr 1;59(1):30-7.

2. Wei Wu,corresponding author Zhiheng Wang, Peisheng Cong, and Tonghua Li. Accurate prediction of protein relative solvent accessibility using a balanced model. 2017 Jan 24. doi: 10.1186/s13040-016-0121-5.

3. Huiling Chen, Huan-Xiang Zhou, Xiaohua Hu, Illhoi Yoo. Classification Comparison of Prediction of Solvent Accessibility From Protein Sequences.

4. Lorenzo Palmieri, Maria Federico, Mauro Leoncini, and Manuela Montangero. A High Performing Tool for Residue Solvent Accessibility Prediction.

5. Manpreet Singh, Parminder Kaur Wadhwa, and Surinder Kaur. Predicting Protein Function using Decision Tree. World Academy of Science, Engineering and Technology 39 2008.

6. Manpreet Singh, Parminder Kaur Wadhwa, and Surinder Kaur. Human Protein Function Prediction using Decision Tree Induction. IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.4, April 2007.

7. https://www.wikipedia.org/