Classification of Protein Information Based on Decision Tree

Ashar Natsheh
Information Technology Department, Palestine Polytechnic University, Wad Alharia, P.O.198,Hebron, Palestine.
ashar@ppu.edu

Yasmin Ta'amrah
Information Technology Department, Palestine Polytechnic University, Wad Alharia, P.O.198,Hebron, Palestine.
yasmeent@ppu.edu

Hashem Tamimi
Information Technology Department, Palestine Polytechnic University, Wad Alharia, P.O.198,Hebron, Palestine.
htamimi@ppu.edu

Abstract— **Using machine learning for automatic classification of protein sequences is a very important problem that obtained a very significant attention. This paper proposes a new automatic classification tool for protein sequences using decision tree learning. We proposed the using of physicochemical properties of amino acids as the attributes of decision tree. The results showed that the proposed method has a good performance in predicting protein classification.**

Keywords-Decision tree; classification; protein sequence; caspase 3.

I. INTRODUCTION

The proteins are the main units of building the cell, and they have a lot of functions that support cell activities. So, the problem of protein function prediction is so important and critical, and considered as one of the most important problems in functional genomics [3].

Classification is the process of assigning data to one of several predefined classes [7].

Decision tree is [7] a machine learning technique that is used for building classification models. DT has a lot of advantages if compared to other learning techniques. It – for example- has the ability for dealing with redundant attributes; also it has a low cost when generating classification models, and many other advantages. Moreover, the model that is inferred from some samples using DT can be used as a general model and gives good results [7] [11].

There are many methods were used to classify proteins [4] [1] [8] [2].

This study presents an approach for predicting caspase 3 substrates cut sites based on the decision tree. The method shows an improvement over existing methods because of using a set of physicochemical properties for each amino acid in each substrate.

In the next section we will introduce some background about decision trees and protein physicochemical properties. Then we will discuss our dataset and method. In section IV we will represent the experiments and results.

II. BACKGROUND

A. Decision Tree Algorithm

Decision tree is a "a predictive modeling technique from the field of machine learning and statistics that builds a simple tree-like structure to model the underlying pattern of data" [11]. Decision tree method has the ability of dealing with qualitative and quantitative data. It can do classification with few computations it also can be easily interpreted.

Decision tree can be considered as a directed tree. Its structure has three types of nodes: The root, which is the starting node, having no incoming edges. Internal splitting nodes, which are none leaf nodes and contains a decision. Leaf nodes, that contains the most appropriate values for a class. Both internal and leaf nodes have exactly one incoming edge.

Decision tree is a good classifier that can simplify a complex decision making process [13]. The decision tree gives the ability to show the results graphically as a tree model. Tree-Building and Tree-Pruning are implemented using CART and C4.5 algorithms in the decision tree procedure [5]. The target of training decision tree from a sample data S is to produce a brief model that is compatible with the training data.

Usually, learning algorithms implements a top-down greedy search algorithm through decision trees, this lead to organizing the tests in a tree. The algorithm gives smaller and better subsets of the data set increasingly giving more nodes by tested under new properties.

Determining which attribute to be chosen next is based on the measure of impurity; which is called the entropy of the sample set for that node, entropy can be calculated according to the following formula [13]:

$$\text{Entropy}(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \quad (1)$$

Where p_i is the fraction of examples in S.

The attribute is as better as how much it decreases the impurity. The decrease in the impurity is called information gain, it can be calculated following this formula [13]:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \quad (2)$$

Where Values (A) represents all possible values for attribute A.

B. Physicochemical Properties of Amino Acids

Amino acid is an organic compound that consists of two kinds of groups: amine group and carboxyl group. A group of amino acids together are forming a protein molecule when linked together by peptide bonds. A set of physicochemical properties are used to classify the 20 amino acids into several groups. These properties were selected from a physicochemical properties database called AAindex which includes about 544 amino acid properties, because of the reputation in this data we selected 50 properties [10].

III. Data Method

A. Dataset

We implemented decision tree on caspase 3 substrates. Our dataset consists of 247 positive (cleaved) peptide data and 247 negative (uncleaved) peptides. Each substrate consists of 14 amino acids [12].

B. Learning

In this step the decision tree is built from the training dataset. The attributes of decision table are the physicochemical properties that each amino acid has in each position of the substrate. So we have 700 values in each row of the table obtained from these properties in addition to one value indicating the classification of the substrate whether it is negative or positive. After that, the decision tree is built using the entropy and gain 1, 2.

C. Evaluation

Different values were used to calculate several measures for classifier evaluation [6], as follows:

False discovery rate:

$$\text{FDR} = \frac{\text{FP}}{\text{FP}+\text{TP}} \quad (3)$$

Specificity:

$$\text{SPC} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (4)$$

Matthews Correlation Coefficient:

$$\text{MCC} = \frac{\text{TP}*\text{TN}-\text{FP}*\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}} \quad (5)$$

Accuracy:

$$\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \quad (6)$$

Precision:

$$\text{PRC} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (7)$$

Where:

TP (True Positive): is the number of correct positive classifications.

FP (False Positive): is the number of incorrect positive classifications.

TN (true negatives): is the number of correct negative classifications.

FN (false negatives): is the number of incorrect negative classifications

All of these measures were calculated using the evaluation values which are resulted from evaluation step.

IV. Experiments and Results

This application for classification the cut sites from the given training set of proteins was implemented in Java. The application uses the Weka version 3.6.9 [9] library to apply the trained classification models and to use them in the classification of the protein.

We have implemented several test cases on the dataset; we used the first 10% of this dataset for testing and the rest 90% for training. Then we used second 10% for testing and so on. Finally we had 10 test cases for evaluation.

Each test case includes 444 instances: 222 of them are negative, and the others are positive. These instances are representing the peptides; each one of them is a substring of 14 characters, and each character has 50 properties representing physicochemical properties, so the whole substring has 700 attributes.

A. Learning

First, we applied the learning method on these instances in order to build the decision tree; the instances are the input of the DT. After learning, the tree has been built; following figure illustrates a part of the tree obtained when applying learning on the instances of the first test case:

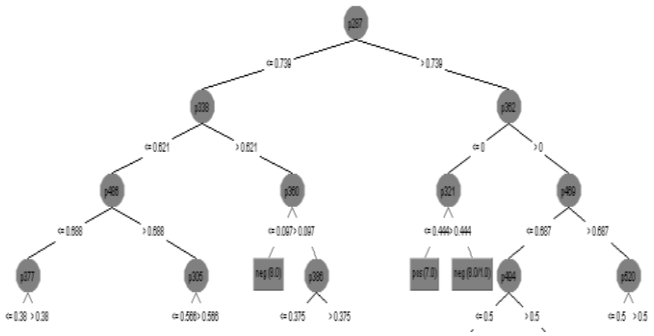

Figure 1: Part of the resulting decision tree

This tree shows that the best attribute to use first is p287, which means that the best position for the character is the 6th position and the best property is "D Relative preference value at N3 " [Richardson-Richardson, 1988].

B. Classification

The classifier of J48 algorithm is provided in the Weka library. It is one version of the C4.5 algorithm [9] a classifier based on decision tree. The training data set are ordered as a sub trees depending on a selected features or attributes that gives the most efficaciously divides the data set. Two methods are followed to obtain the classification results. First, is

information entropy which calculates the impurity measure, and then splitting the tree with the best split, this method is still iterating until the tree cannot be splitted any more. Second method is using pruning in order to improve the accuracy of the prediction by taking off the useless tree nodes. The final results of the classification are assigned to the leaf nodes indicating the class.

C. Evaluation

The third step is the evaluation of the data appending on the model giving different measures and statistics.

Final step is testing the efficiency of the model (the tree), by applying classification method on the 10% of the data set that are reserved for this method.

The first test case gave a result of 78% accuracy in the classification method.

When testing the negative data, 5 of the instances were not classified correctly, while 20 of them obtained the right classification. For the positive test, 6 instances were incorrect while 19 were classified correctly. The average of correct instances classification obtained via several test cases was 76%.

The evaluation of the classifier with quality measures indicated that the proposed method has a good performance in predicting caspase 3 cleavage sites.

CONCLUSION

We have developed a decision tree tool for reading a number of proteins sequences, predicting their caspase cut sites and outputting the positives sequences. We show that our method is a worthy tool in identifying novel caspase target proteins from proteomics experiments. The importance of this work is the using of physicochemical properties of amino acids.

FUTURE WORKS

For the physicochemical properties method we used 50 of the most important physicochemical properties, and we need to use additional amino acids physicochemical properties, but the decision tree needs more work.

REFERENCES

[1] Cathy H. Wu, Hongzhan Huang, Lai-Su L. Yeh and Winona C. Barker, Protein family classification and functional annotation, Computational Biology and Chemistry 27, 2002.

[2] Christina Backes, Jan Kuentzer, Hans-Peter Lenhof, Nicole Comtesse and Eckart Meese, "GraBCas: bioinformatics tool for score-based prediction of Caspase- and Granzyme Bcleavage sites in protein sequences," Nucleic Acids Research, 2005, Vol. 33.

[3] Eduardo P. Costa, Ana C. Lorena, Andr´e C. P. L. F. Carvalho, Alex A. Freitas3, and Nicholas Holden, "Comparing Several Approaches for Hierarchical Classification of Proteins with Decision Trees," Sao Carlos-SP, Brazil.

[4] Jason Weston, Christina Leslie, Dengyong Zhou, Andre Elisseeff and William Stafford Noble , "Semi-supervised protein classification using cluster kernels,".

[5] Lan H.Witten and Eibe Frenk. "Data mining practical Machine learning tools and techniques," second edition, 2005.

[6] Mirva Piippo, Niina Lietzén, Olli S Nevalainen, Jussi Salmi and Tuula A Nyman, "Pripper: prediction of caspase cleavage sites from whole proteomes," *BMC Bioinformatics* 2010.

[7] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, "Introduction to Data Mining," chapter 4, March 25, 2006.

[8] R. H. M. Garay-Malpartida, J. M. Occhiucci, J. Alves and J. E. Belizário, "CaSPredictor: a new computer-based tool for caspase substrate prediction," Vol. 21 Suppl. 1 2005, pages i169–i176 Vol. 21 Suppl. 1 2005, pages i169–i176. [pc 3]

[9] Remco R. Bouckaert, Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald and David Scuse. "WEKA Manual for Version 3-6-9," January 21, 2013.

[10] Shuichi Kawashima, Piotr Pokarowski, Maria Pokarowska, Andrzej Kolinski, Toshiaki Katayama, and Minoru Kanehisa. Aaindex: amino acid index database, progress report 2008. Nucleic Acids Res, 36(Database issue):D202–D205, Jan 2008.

[11] Tejaswini Pawar and Prof. Snehal Kamalapur. "A Survey on Privacy Preserving Decision Tree Classifier," International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 6, November- December 2012, pp.843-847.

[12] The Caspase Substrate database Homepage. [http://bioinf.gen.tcd.ie/ casbah/].

[13] Wijai Boonyanusith and Phongchai Jittamai. "Blood Dono Classification Using Neural Network and Decision Tree Techniques," Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I.