Harnessing Power of Decision Tree Approach for HPF Prediction using SIPINA and See5

Sunny Sharma [#1], ***Amritpal Singh*** [#2], **Dr. Rajinder Singh** [#3]

[#1] *Research Scholar, Department of Computer Science, Guru Nanak Dev University, Amritsar, India*

[#2] *Research Scholar, Department of Computer Science, Guru Nanak Dev University, Amritsar, India*

[#3] *Professor& Head, Department of Computer Science, Guru Nanak Dev University, Amritsar, India*

Abstract— Drug discovery process, Disease detection and Prediction of molecular class are the area of great significance for carrying out research. In past few decades some precise approaches were used to enhance the accuracy of Human protein Function (HPF) prediction. This research study is primarily concentrated on such approach of HPF prediction with sequence derived features (SDF) using decision trees and there variants implemented using C5 and C4.5 algorithms like See5 and SIPINA. More sequence derived features were identified and incorporated. The training data was improved with these incorporated features. The Sequence data was evolved from HPRD (Human protein reference database) in terms of number of sequences and the features used to extract the relation towards a specific class which enhancing power of training data. Multiple techniques were examined for accuracy in prediction and a widespread comparison was done amongst them incorporating with previous research results, and prescribed the overall accuracy of See5 with 64% and SIPINA with 88%.

Keywords— *HPF, C5, C4.5, See5, Decision Tree, SDF,* SIPINA

I. INTRODUCTION

A time tried and tested approach of prediction is decision tree based prediction. It is a white-box technique which clearly illustrates the sequence of computations involved at each and every stage. This plus point enables its usage by computational experts even without much knowledge of the concerned domain. Likewise, it enables an expert from the concerned domain to critically examine the steps followed by a computational expert. So it bridges the gap between technical know-how and domain expertise. Decision tree comprises of nodes and edges depicting various functionalities at different levels of computations. A decision tree clearly illustrates the required results or outputs amongst various outcome possibilities. It clearly defines the problem structure and its interpretations in a hierarchical way which is much easier to comprehend. As the model has a unique ability of taking into account various input parameters and reaching a goal.

II. DECISION TREE FOR CLASS DETECTION

Decision trees approach is a very potent methodology of supervised learning. A set of classified data is provided as input and a tree that signifies an orientation diagram having each of the leaf nodes as a class i.e. decision and each internal node indicates a test that is obtained as an output. Decision of relation to a class of data is indicated by each of the leaf confirming to the entire tests path from the root node to that of the leaf node.

III. INTRODUCTION TO SIPINA

The crucial problem in data mining is to handle large databases and efficacy of SIPINA is to handle large databases and discover the hidden information in large databases. SIPINA is a not only a data mining tool it has machine learning capabilities also. But specialty of SIPINA is intended to decision trees induction or we can say classification trees, use supervised learning. it is free for all kind of activities. SIPINA is circulated on the web since 1995, SIPINA is classification tree developer incorporated with dedicated classification trees algorithms like ID3, GID3, ASSISTANT 86, CHAID, C4.5, One Vs All Decision Tree etc. it also has some other mining capability through Rule Induction, Neural Network, Discriminant Analysis, Decision List etc. we can say some supervised methods are also accessible e.g. K-NN, Multilayer perceptron, Naive Bayes, etc. we can use any one of them one at a time. It corresponds to an algorithm for the induction of decision graphs. Experiment is done on Human protein data bank. Small sample of data set contains five diverse classes with 23 parameters or attributes [11].

IV. C5 ALGORITHM IMPLEMENTATION OF SEE5

Quinlan's C5.0 algorithm is widely used for classification process. Algorithm primarily focuses on constructing a decision with the identification of most important attributes from the supplied/identified data-set. Once the attribute is finalized from current node, corresponding child nodes are then generated. There after best attribute of a node can be selected. There are few options present over See5 tool like, Boosting is to generate several classifiers (decision trees or rule-sets) instead of one. On classifying a new case, each classifier supports its predicted class and then the support is evaluated to determine the final class. In the first step, a single decision tree or rule-set is constructed as before from the training data. This classifier will usually make mistakes on some cases,

like here the first decision tree, gives the wrong class for 14 cases in sequence data. Other classifier is constructed giving more consideration to the cases. Thus the classifier will provides results variation from the earlier classifier. Errors induced are again rectified by another classifier. It continues for defined iterations/trials and halts once extremely correct classifies is achieved [3], [4].

Winnowing is a mechanism to separate the useful attributes from useless attributes. It provides option to select among the predictors and have an edge to create a suitable decision-tree. However, it's time intensive task and primarily suitable for bigger application domain. [3], [10].

In Advanced pruning technique a massive tree is first allowed to grow to fit the data closely after that it's pruned i.e. error causing segments are removed. Every sub-tree undergoes pruning then replacement by a leaf or sub branch is decided and then a global stage evaluates performance of the tree as a unit. [3], [10].

V. Literature Survey

Jensen, L. et al. (2002) focused on developing fully sequence-based method that recognizes and combines important features for the purpose of assigning proteins of unknown function to respective classes and enzyme classification. A number of functional features that are more appropriately related to the linear sequence of amino acids may benefit the strategies for the elucidation of protein function, and hence quite simple to predict, than protein structure. Identified Attributes include features associated with post-translational modifications and protein sorting; also include simpler aspects such as the length, composition of the polypeptide chain and isoelectric point [6]. Friedberg, I. (2006) showed that not only is the volume and diversity of pure sequence and structure data is increasing and resulting to a unequal growth in the number of uncharacterized gene products. Consequently, established methods of gene and protein annotation, such as homology-based transfer, are annotating less data and in many cases are amplifying existing erroneous annotation. Also functional annotation is desired which is standardized and machine readable for the requirement of prediction programs implementation on larger workflows. Subjective and contextual definition of protein function is cumbersome in nature. The need to assess the quality of function predictors needs to be stressed upon [4]. Singh, M. et al. (2007) exponential increase in protein data was suggested to solve the problem; drug discoverers need efficient machine learning techniques to predict the functions of proteins which are responsible for various diseases in human body. Decision tree induction methodology used in C4.5 for the selection of best attribute involves the entropy calculation. For the discrete same test data, the correctness of the new HPF (Human Protein Function) predictor was 72% and that of the existing prediction methodology was 44% [8]. Singh, M. et al. (2011) presented cluster analysis as a form of unsupervised learning and cluster analysis is implemented for human protein class prediction. The data is accessed from Human Protein Reference Database (HPRD) which is related to human protein. The sequences related to ten molecular classes are obtained using HPRD. Five amino acid sequences are obtained for each of the molecular class. SDFs (Sequence derived Features) are extracted for each sequence by using various web based tools. On the basis of values of input SDFs and by considering priority of each of the SDF, clusters of the data available in the adjacency matrix are generated. Then those clusters are backtracked to predict the class of the entered sequence [7]. Rule-based classification is preferred because it is easy to comprehend and the reason lies in examining and validation of every rule individually without bothering about its holistic impact. See5 (Implementing C5) is an excellent tool when performance is taken into account. Decision trees are generated and they are of great use when quick construction of the classifiers is required. [10]. Arditi, D. et al. (2005) examined construction litigation application of See5. A boosted decision-tree system approach was incorporated to predict the results of construction-litigation domain. Same data-sets as used in previous prediction related examination conducted with ANN's earlier and case-based reasoning afterwards were included in the research, augmented by an additional cases that were filed in 1990–2000. All cases were extracted from the Westlaw on-line service. Boosted decision trees provided a superior prediction accuracy of 90%. [2]. Wei-Feng, H. et al. (2011) demonstrated that the link between the synthetic features and the types of final product are very important for the material's rational synthesis. A prediction mechanism was proposed that was C5centric and combined with a feature selection. Classification accuracy and a receiver operating characteristic (ROC) curve determined the performance credential for the proposed methodology. Highest area under the ROC curve (90%) and the classification accuracy (88.18%) was achieved in results for the decision tree model containing 8 input attributes and some important rules with high confidence degrees were extracted from the model [3].

VI. Implementation on See5 and SIPINA

On the basis of different sequences of human protein the on See5 the C5 algorithm implementation predicts the molecular class and on SIPINA C4.5 algorithm implementation predicts the molecular class. In the sample data 15 Protein classes, with 70 protein sequences is taken and each of them is having 25 attributes or features on See5 as well as on SIPINA 5 Protein classes, with 25 protein sequences is taken and each of them is having 23 attributes or features [9], [10].

6.1 Cross-Validation

The real prediction correctness of a classifier can be evaluated by sampling i.e. using different test files rather than relying only on training data. So cross validation is done using unseen data as well and enhances accuracy in the prediction process.

VII. RESULTS & DISCUSSIONS

On See5 the dataset containing 70 sequences and 25 features was examined and the correctness of various techniques is depicted in fig. 1. The C5 algorithm with winnowing and advance pruning option provides the maximum accuracy of 45%. If the same number of elements are taken as that of [8], the accuracy comes out to be 64%. But in SIPINA with 25 sequences and 23 features was examined and the accuracy id achieved in the prediction is 88%.

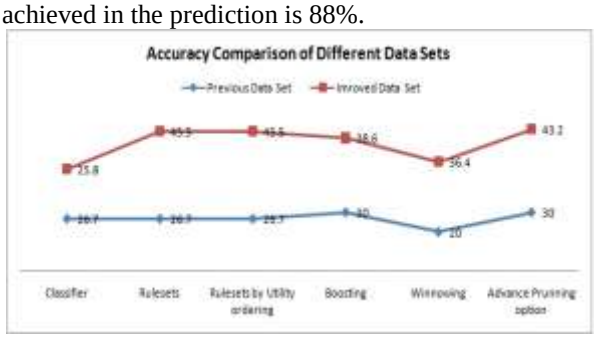

Fig. 1: Accuracy Comparison in See5

The goodness of split of all the attributes is obtained from the data using SIPINA and the attribute expaa outperforms. The correlation of all the attributes with goodness of fit is also calculated which is shown on fig. 2 with attribute acceptance.

	Goodness of split	Correlation	Accept or Reject
expaa	0.85409515	0.4351	
prob	0.84683548	0.4306	
t	0.77237357	0.3810	
s	0.77237357	0.3810	
mean	0.64083826	0.3264	
d	0.60421367	0.3016	
ser	0.60000000	0.3000	
thr	0.60000000	0.3000	
npos	0.60000000	0.3000	
nneg	0.60000000	0.3000	
exc1	0.56228545	0.3052	
exc2	0.56228545	0.3052	
instabilityindex	0.51715523	0.1739	
predhel	0.51195690	0.2727	
tyr	0.46622948	0.2024	
gravy	0.46597605	0.2426	
aliphaticindex	0.32206587	0.1795	

Fig. 2: Goodness of split of attributes using SIPINA

The splitting suggestion for all the classes using expaa attribute is shown which prescribe the critical value of 4.94 for the classification of HPF classes. Using this value the decision tree is shown in Fig. 5 and the splitting suggestion is shown in Fig. 3.

Split suggestion for "expaa"

	< 4.94	>=4.94
defensin.	1	4
voltagegate	0	5
dnarepairpr	5	0
heatshockp	5	0
cellsurfacer	0	5

Fig. 3: Splitting suggestion attributes using SIPINA

And the Decision trees obtained with See5and SIPINA are shown as follows:

- See5 (shown in Figure: 4)
- SIPINA (shown in Figure: 5)

VIII. CONCLUSION

Present work focus on harnessing the power of decision tree approach for HPF prediction using Sipina and See5 and also demonstrate the impact of choosing the right training data. The detailed analysis shows that increasing number of features (5 features) of HPF data increases the accuracy of prediction process (about 16%) in See5. but does not necessarily involves the participation of all parameters in decision making process. Similarly the experiments are carried out with SIPINA which shows the overall accuracy of 88% through confusion matrix. Some parameters were more dominant than others like GRAVY 13%, Solubility 8%, Thr 4% etc. in See5 and expaa with critical splitting value of 4.94. Hence they decide the course of prediction. Activities like advanced pruning and winnowing (17 attributes winnowed) help in minimizing the computation time and also help in reaching the most important parameters involved in prediction process. ExpAA came out as most important parameter after winnowing in both See5 and SIPINA.

In future more features can be extracted on single sequence and their relative impact on prediction process can be examined hence it will lead to greater precision in the HPF identification process.

ACKNOWLEDGMENTS

Our thanks to Department of Computer Science & Engineering, Guru Nanak Dev University, Amritsar, India for providing us all the required resources necessary to study and implement the solution to this problem. We also want to thank all the authors of referenced papers which guided us all the time as base for this study.

REFERENCES

[1] B. Bergeron, "Bioinformatics Computing", pp 257-270, 2002.

[2] D. Arditi and T. Pulket, "Predicting the outcome of construction litigation using boosted decision trees ", *Journal of Computing in Civil Engineering*, vol. 19, no. 4, pp 387–393, 2005.

[3] H. Wei-Feng, G. Na, Y. Yan, L. Ji-Yang, Y. Ji-Hong, "Decision Trees Com-bined with Feature Selection for the Rational Synthesis of Aluminophos-phate AlPO4-5", *National Natural Science Foundation of China*, vol 27, no.9, pp 2111-2117, 2011.

[4] I. Friedberg, "Automated Protein Function Prediction- the Genomic Chal-lenge", *Briefings in Bioinformatics*, vol 7, no.3, pp 225-242.

[5] J. Han and M. Kamber, "Data Mining Concepts and Techniques", *MorganKaufmann Publishers, USA* pp 279-322, 2003.

[6] L.J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames C. Kesmir, H. Nielsen, H.H. Stærfeldt, K. Rapacki, C. Workman C.A.F. Andersen, S. Knudsen, A. Krogh, A.Valencia and S. Brunak , "Prediction of Human Protein Function from Post-Translational Modifications and Localization Features ", *Journal of Molecular Biology*, vol. 319, issue 5,pp 1257-1265, 2002.

[7] M. Singh, G. Singh, "Cluster Analysis Technique based on Bipartite Graph for Human Protein Class Prediction", *International Journal of Computer Applications (0975 – 8887)*, vol. 20, no.3, pp. 22-27, 2011.

[8] M. Singh, P. K. Wadhwa and P. S. Sandhu , " Human Protein Function Prediction using Decision Tree Induction ", *IJCSNS International Journal of Computer Science and Network Security*, vol. 7, no.4, pp. 92-98, 2007.

[9] www.hprd.org.

[10] http://rulequest.com/see5-info.html.

[11] http://eric.univ-lyon2.fr/~ricco/sipina.html

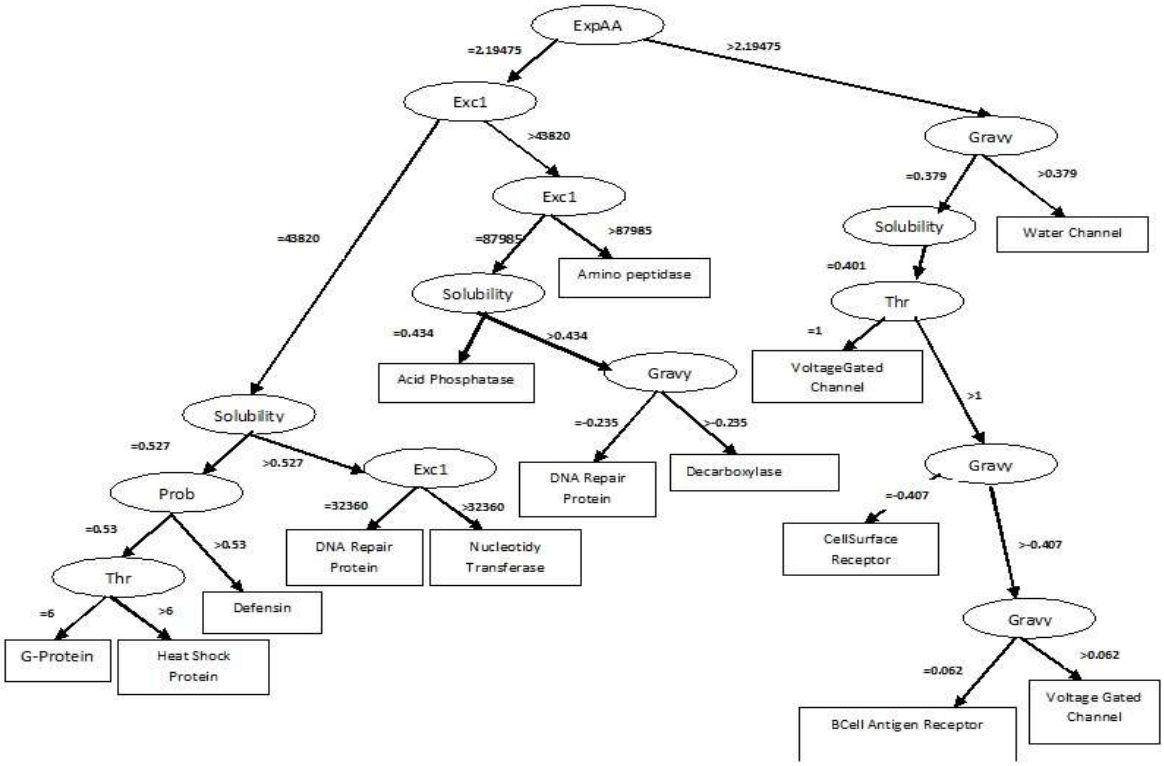

Fig. 4: Decision Tree Implementation with See5

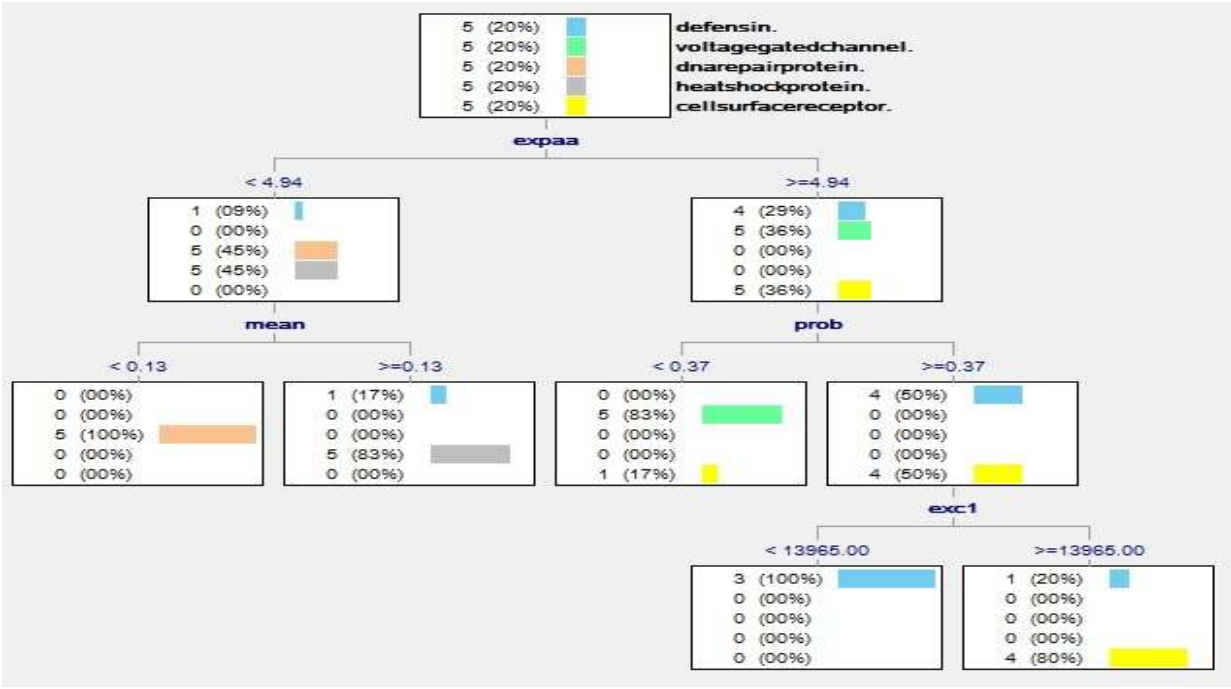

Fig. 5: Decision Tree Implementation on SIPINA