Human Protein Function Prediction using Decision Tree Induction

Manpreet Singh[†], **Parminder Kaur Wadhwa**[††] **and Parvinder Singh Sandhu**[†††]

Deptt. Of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana, Punjab, INDIA

Summary

To overcome the problem of exponentially increasing protein data, drug discoverers need efficient machine learning techniques to predict the functions of proteins which are responsible for various diseases in human body. The existing decision tree induction methodology C4.5 uses the entropy calculation for best attribute selection. The proposed method develops a new decision tree induction technique in which uncertainty measure is used for best attribute selection. This is based on the study of priority based packages of SDFs (Sequence Derived Features). The present research work results the creation of better decision tree in terms of depth than the existing C4.5 technique. The tree with greater depth ensures more number of tests before functional class assignment and thus results in more accurate predictions than the existing prediction technique. For the same test data, the percentage accuracy of the new HPF (Human Protein Function) predictor is 72% and that of the existing prediction technique is 44%.

Key words:

Decision Tree Classifier (DTC), Sequence Derived Features (SDFs), entropy, Uncertainty measure.

1. Introduction

The importance of human protein function prediction lies in the sensitive procedure of drug development. Drug development involves two major components – drug-discovery and testing [9]. The testing process involves preclinical and clinical trials. The computational methods are not generally subjected to produce significant enhancement in testing processes of drugs. But in the discovery process the efficient computational methods are needed. The drug discovery process is labor intensive and expensive. The process of drug discovery, involves the prediction of protein function based upon existing facts. Sophisticated data mining models are needed for protein function prediction. Bioinformatics promises to reduce the labor, time as well as cost associated with this process.

1.1 Protein Function Prediction Techniques

The computational techniques for predicting the structure and functions of unknown proteins are as follows:

A. The QM/MM scheme i.e. the Quantum Mechanical/Molecular Mechanical scheme is used by software named GAMESS (General Atomic and Molecular Electronic Structure System) to predict an unknown protein. It requires a large computer memory to perform mathematical calculations and it runs on Linux operating system.

B. Software named as SWISS-Model is available for automated building of the theoretical structural models of a given protein (amino-acid sequence) based on the known proteins' structures.

C. Classifiers, for example, neural networks, decision trees etc. learn classification rules from the given training data which are used to predict functions of unknown proteins.

1.2 Decision Tree Classifier (DTC)

A decision tree is a classifier which can work efficiently over large volumes of data. It is a flow-chart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and leaf nodes represent classes or class distributions. The top-most node in a tree is the root node [2].

In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample.

2. Problem Statement and Solution Approach

2.1 Attribute Selection in DTC based on Information

According to Information Theory (by Shannon), if the total information to be transmitted is divided into certain and uncertain, and lesser number of bits are assigned to a sequence of certain information than uncertain information, then on an average, lesser number of bits are needed to be transmitted over the communication channel [1].

Let the variable *x* range over the values to be encoded, and let *P(x)* denote the probability of that value occurring. Then, according to Information Theory, the expected number of bits required to encode one value is the weighted average of the number of bits required to encode each possible value, where the weight is the probability of that value [1]:

$$\sum_{x} P(x)*-\log_2 P(x) \tag{1}$$

The DTC (Decision Tree Classifier) methodology involves entropy calculation. Entropy is the expected information based on the partitioning into subsets by an attribute. The smaller the entropy value, the greater is the purity of the subset partitions.

The information gain measure is used to select the test attribute at each node in the tree [2] [19]. The attribute with the highest information gain (or greatest entropy reduction) is chosen as the test attribute for the current node. Let *S* be a set consisting of s data samples. Suppose the class label attribute has *m* distinct values defining m distinct classes: C_i (for i = 1, . . . , m). Let s_i be the number of samples of *S* in class $C_{i.}$ The expected information needed to classify a given sample is given by [2]:

$$I(s_1, s_2, ..., s_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \tag{2}$$

Where, p_i is the probability that an arbitrary sample belongs to Class C_i and is estimated by s_i/s. Let attribute *A* have *v* distinct values, {a_1, a_2, . . . ,a_v}.Attribute *A* can be used to partition *S* into *v* subsets, {S_1, S_2, . . ., S_v}, where S_j contains those samples in *S* that have value a_j of *A*.

If *A* were selected as the test attribute (i.e., the best attribute for splitting), then these subsets would correspond to the branches grown from the node containing the set *S*. Let s_{ij} be the number of samples of class C_i in a subset $S_{j.}$ The entropy, or expected information based on the partitioning into subsets by *A*, is given by [2]:

$$E(A) = \sum_{j=1}^{v}((s_{1j} + ... + s_{mj}) / s) I(s_{1j}, ... s_{mj}) \tag{3}$$

The term $((s_{1j} + . . . + s_{mj})/s)$ acts as the weight of the j^{th} subset and is the number of samples in the subset (i.e. having value a_j of *A*) divided by the total number of samples in *S*. The smaller the entropy value, the greater the purity of the subset partitions. The encoding information that would be gained by branching on *A* is [2]:

$$Gain(A) = I(s_1, s_2, ..., s_m) - E(A) \tag{4}$$

The *Gain(A)* is the expected reduction in entropy caused by knowing the value of attribute *A*.

The algorithm computes the information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given set *S*. A node is created and labeled with the attribute, branches are created for each value of the attribute, and the samples are partitioned accordingly [12].

2. HPF Prediction using SDFs

Jensen used sequence derived features to predict HPF. The idea was to integrate all protein features in order to predict protein function. The author developed the data mining model for protein function prediction using neural networks as classifier. The method used by the author includes the extraction of SDFs from a given set of amino-acid (protein) sequences using various web-based bioinformatics' tools. For example, ExPASy ProtParam tool is used to obtain the sequence-derived feature called Extinction Coefficient which is a protein parameter that is commonly used in the laboratory for determining the protein concentration in a solution by spectrophotometry. It describes to what extent light is absorbed by the protein and depends upon the protein size and composition as well as the wavelength of the light. For a wavelength of 280 nm, the Extinction Cofficient of a protein can be calculated from the number of tryptophans (n_{Trp}), tyrosines(n_{Tyr}) and cystines(n_{Cys}) in the protein [8] as shown below:

$$\varepsilon_{protein} = \eta_{Trp}\varepsilon_{Trp} + \eta_{Tyr}\varepsilon_{Tyr} + \eta_{Cys}\varepsilon_{Cys} \tag{5}$$

Where, ε_{Trp}, ε_{Tyr} and ε_{Cys} are the extinction coefficients of the individual amino-acid residues. This calculation is performed by the ExPASy ProtParam tool. Similarly, other sequence-derived features are abtained from the ExPASy ProtParam tool and others as shown in Table 1.

Table 1: SDFs Obtained From Various Web Tools

Tool used	*SDF Obtained*
ExPASy ProtParam	Extinction Coefficient Hydrophobicity No. of negatively charged residues No. of positively charged residues
NetNglyc	N-glycosylation sites
NetOglyc	O-glycosylation sites
NetPhos	Sr and Thr phosphorylation Tyr phosphorylation
PSI-Pred	Secondary Structure
PSORT	Subcellular Location
SignalP	Signal Peptide
TMHMM	Transmembrane Helices

Thus the following equation is obtained:

sequence→ sequence-derived features (6)
[using various Bioinformatics Tools]

For the same set of protein sequences, the protein functions are obtained, i.e.

sequence→ predicted protein features (7)

By combining the Eq. 6 and Eq. 7 :

sequence→ A List of sequence-derived features (8)
+
predicted protein functions

The existing technique of HPF prediction does not involve the consideration of the factor of dominancy of SDFs for particular functional class. The proposed method is to use priority based packages of SDFs so that decision tree may be created by their depth exploration rather than exclusion.

3. Solution Methodology

3.1 Data Collection and Preprocessing

The actual data related to human protein is accessed from Human Protein Reference Database (HPRD). The HPRD represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks and disease association for each protein in the human proteome. All the information in HPRD has been manually extracted from the literature by expert biologists who read, interpret and analyze the published data. It includes approximately 162 classes of protein functions. The database provides information about protein function under the heading ‘molecular class’ covering all the major protein function categories.

From HPRD, the sequences related to five molecular classes are obtained. These are: Defensin (Def),Cell Surface Receptor (CSR), DNA Repair Protein (DRP), Heat Shock Protein (HSP) and Voltage Gated Channel (VGC). Various web-based tools are then used to derive SDFs from these sequences. The SDFs are preprocessed by placing their values in particular value ranges to make them suitable for input to classifier.

3.2 Packages of SDFs

The actual data related to human protein is accessed from Human Protein Reference Database for creating packages of SDFs:

- The frequencies of values of SDFs are studied for each functional class. If a particular value of SDF repeats very highly for a particular molecular class, then it is considered as dominant for that class.
- On the basis of the dominancy, packages of SDFs are obtained.
- The packages of SDFs obtained are shown in Table 2. These packages are used to create various decision trees.

Table 2: Packages of SDFs Obtained

Package of Features	Pack 1	Pack 2	Pack 3	Pack 4
ExPASy ProtParam:				
Nneg	✗	✗	✓	✓
Npos	✗	✗	✓	✓
Exc1	✗	✗	✗	✓
Exc2	✗	✗	✗	✓
Instability Index	✗	✗	✓	✓
Aliphatic Index	✗	✓	✓	✓
GRAVY	✗	✗	✓	✓
NetOGlyc:				
T	✗	✗	✓	✓
S	✓	✓	✓	✓
NetPhos:				
Ser	✗	✓	✓	✓
Thr	✓	✓	✓	✓
Tyr	✓	✓	✓	✓
SignalP:				
mean S	✗	✗	✓	✓
D	✗	✗	✓	✓
Probability	✗	✓	✓	✓
TMHMM:				
ExpAA	✓	✓	✓	✓
PredHel	✓	✓	✓	✓

The implementation of the data mining model that creates decision trees on the basis of packages of SDF chosen, demonstrates that the use of more dominant SDFs in decision tree creation affects the depth of tree. The decision tree with maximum depth of eight is obtained by using this technique.

But, this technique involves the drawback of the overhead of creating package of SDFs by studying their dominancy for a particular molecular class.

3.3 New Prediction Technique

The new prediction technique incorporates the effect of choosing dominant SDFs for decision tree creation during entropy (or uncertainty) calculation itself. It overcomes the limitation of the model involving packages of SDFs as it does not involve the overhead of creating packages of SDFs. The technique does not encode the information in terms of bits, as it is not required in this application.
The technique considers following factors for measuring uncertainty:

A) Uncertainty due to subset creation (S_u): The ratio of the number of samples of all classes having value j of attribute A to the total number of samples in S, indicates the uncertainty due to subset creation [2].

$$S_u = \sum_{j=1}^{v}((s_{1j} + s_{2j} + ... + s_{mj}) / s) \tag{9}$$

Where, S_u is uncertainty due to subset creation, j is a particular value of attribute A, v is the total number of values of attribute A and m is the total number of classes
This factor indicates the entropy (or uncertainty) caused due to the creation of subset for a value of an attribute. If S_u is high, uncertainty is high, i.e.

$$U \,\alpha\, S_u \tag{10}$$

Where, U indicates uncertainty measure.

B) Specificity (or certainty) of a value of an attribute for a particular class (S_p): The ratio of the number of samples of class having value j of attribute A to the total number of samples of that particular class, indicates the specificity (or dominancy) of a value for the class.

$$S_p = \sum_{j=1}^{v} \sum_{i=1}^{m} (s_{ij} / s_i) \tag{11}$$

Where, S_p is specificity (or certainty) of an attribute-value for a particular class, i is a particular molecular class, j is a particular value of attribute A, v is the total number of values of attribute A and m is the total number of classes.
This factor indicates certainty of a particular functional class for an attribute value. If S_p is high, certainty is high and thus uncertainty is low, i.e.

$$U \,\alpha\, 1 / S_p \tag{12}$$

Combining equations (10) and (12), we get:

$$U \,\alpha\, S_u / S_p \tag{13}$$

Or

$$U = \frac{S_u}{k + S_p} \tag{14}$$

Where, k is a constant.
For k = 1, S_p with value greater than zero, (i.e. $S_p > 0$) can only contribute to the calculation of uncertainty measure. Thus:

$$U = \frac{S_u}{1 + S_p} \tag{15}$$

i.e.

$$U = \frac{\sum_{j=1}^{v}((s_{1j}+........+s_{mj}) / s)}{1 + \sum_{j=1}^{v}\sum_{i=1}^{m}(s_{ij} / s_i)} \tag{16}$$

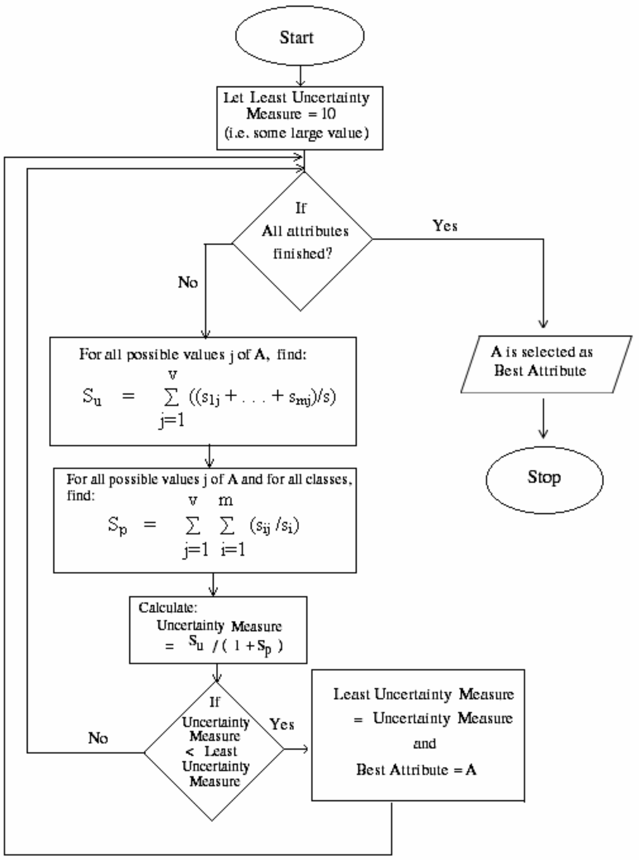
Start
Let Least Uncertainty Measure = 10 (i.e. some large value)
If All attributes finished?
Yes
No
For all possible values j of A, find: $S_u = \sum_{j=1}^{v} ((s_{1j} + . . . + s_{mj})/s)$
For all possible values j of A and for all classes, find: $S_p = \sum_{j=1}^{v} \sum_{i=1}^{m} (s_{ij} / s_i)$
Calculate: Uncertainty Measure = S_u / (1 + S_p)
If Uncertainty Measure < Least Uncertainty Measure
No
Yes
Least Uncertainty Measure = Uncertainty Measure and Best Attribute = A
A is selected as Best Attribute
Stop

Fig. 1 Flowchart for Best Attribute Selection by New Prediction Technique

The attribute with the least uncertainty measure is chosen as the best attribute by the new prediction technique during decision tree creation. Fig. 1 shows the flowchart for best attribute selection on the basis of least uncertainty measure by new prediction technique.

3.4 Materials and Methods Used

The training data consists of labeled feature vector related to five molecular classes of HPRD. Five sequences of each molecular class is accessed from HPRD. Through web-based bioinformatics' tools, SDFs are obtained from these sequences and are processed in a form suitable for input to classifier by placing them in particular value ranges. The processed SDFs and the molecular classes are then combined to form labeled feature vector (or training data).
The classifier for protein function prediction is the modified decision tree induction technique that uses uncertainty measure for the best attribute selection. The attribute with the least uncertainty measure is chosen as the best attribute.

3. Results and Discussion

Fig. 2 demonstrates the use of HPF predictor by drug discoverer. A decision tree is created by HPF predictor by using training data (i.e. processed SDFs and known functional classes) by using new decision tree induction technique. The test data is used to compute percentage accuracy of the decision tree created.
For the same training data, the existing technique creates decision tree with depth of two nodes (as shown in Fig. 3) while the new prediction technique creates decision tree with depth of thirteen nodes **(**as shown in Fig. 4). The large depth of the tree has led to the consideration of more number of tests before functional class assignment and thus has resulted in more accurate predictions. For the same test data, the percentage accuracy of the new HPF predictor is 72% and that of the existing prediction technique is 44%.
The implementation of the model demonstrates the influence of S_p (i.e. specificity or certainty of an attribute-value for a particular class) on the tree creation. A high value of S_p lowers the value of uncertainty measure (as *uncertainty measure* α 1/ S_p) and thus contributes in the best attribute selection for tree creation.

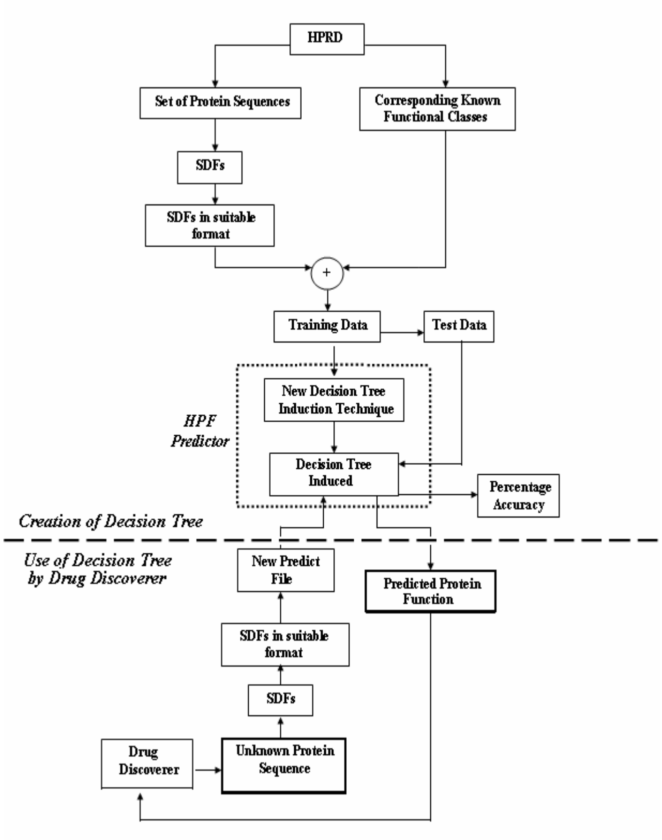

Fig. 2 Use of HPF Predictor by Drug Discoverer

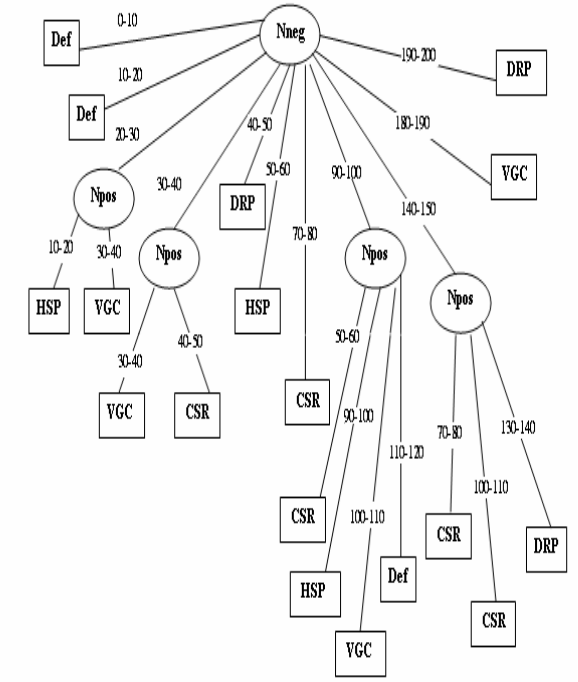

Fig. 3 Decision Tree Created by Existing Prediction Technique

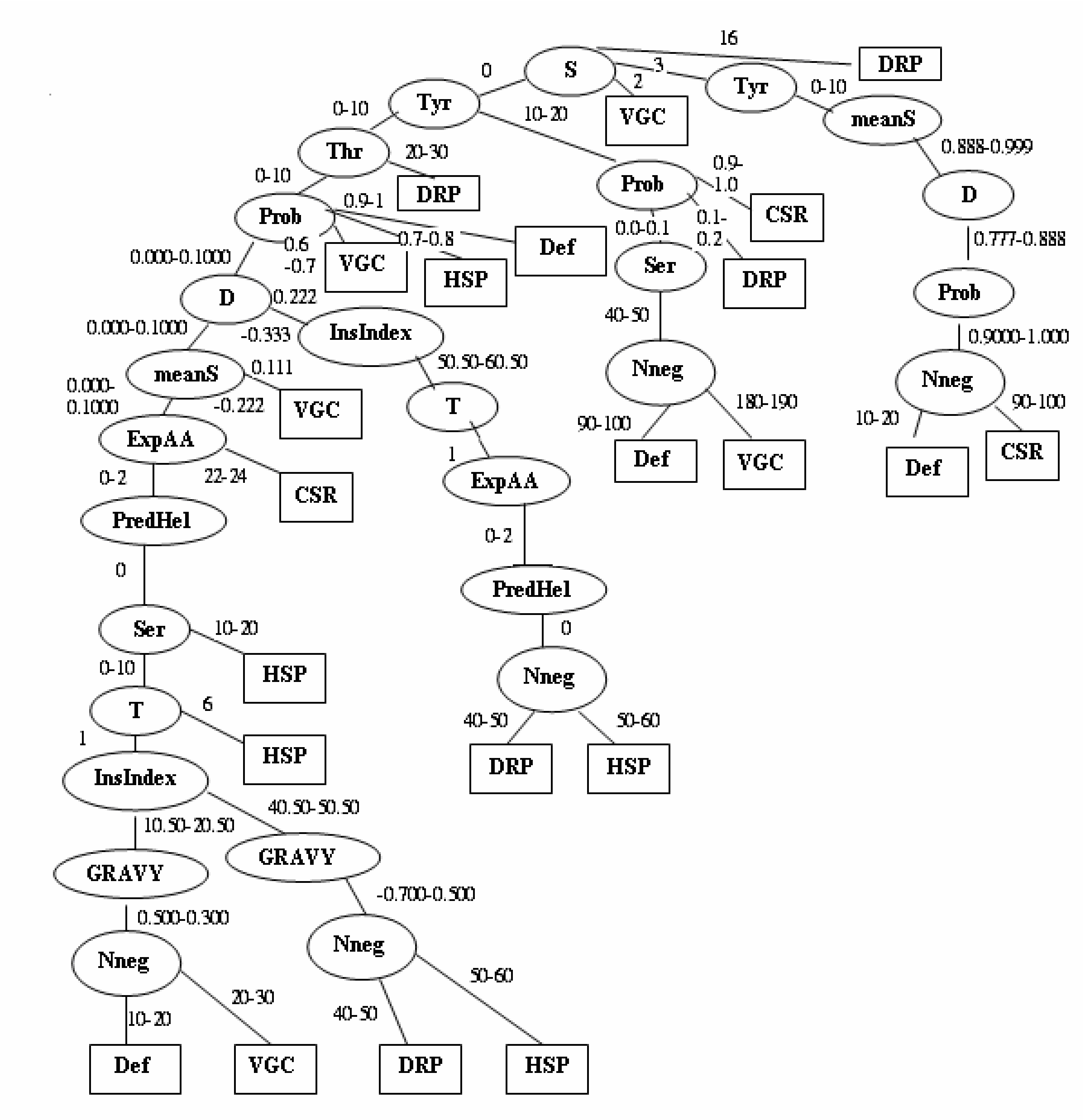

Fig. 4 Decision Tree Created by New Prediction Technique (Involving thirteen nodes)

The implementation of the model shows that the computation of uncertainty measure by the new prediction technique is superior than the entropy calculation in the existing technique. The new prediction technique provides the decision tree with better quality (in terms of depth) than the existing methodology. The depth of the decision tree is directly proportional to the S_p (i.e. the specificity of a value of an attribute for a particular class).

$$\text{Depth of Decision Tree} \quad \alpha \quad S_p \qquad (17)$$

Greater the value of S_p for a particular value of an attribute, more is the chance of developing a branch from that value to a functional class. The new decision tree induction technique provides tree with greater depth than the existing methodology due to the consideration of S_p. The greater depth of the decision tree provides greater number of tests before functional class assignment and hence provides more accurate prediction results.

4. Conclusion

The data mining model for HPF prediction provides better classification rules for the same training data than the existing technique. The model creates better-quality decision tree (in terms of depth) and hence ensures more accurate predictions than the existing methodology. Drug discoverers can easily use the model for predicting functions of proteins that are responsible for various diseases in human body. The steps required for the use of new prediction technique by the drug discoverer are clearly demonstrated. There is large scope for application of the new prediction technique in drug discovery process due to its better quality and clear representation of the learned classification rules.

References

[1] D.S. Touretzky "Basics of Information Theory. Computer Science Department", *Carnegie Mellon University, Pittsburgh*, PA 15213: 2004.

[2] J. Han, and M. Kamber *Data Mining: Concepts and Techniques,* Morgan Kaufmann Publishers. 2004.

[3] H. Almuallim, et al. "Development and Applications of Decision Trees", *Information and Computer Science Department*, 11, 2003, pp. 1374-1379.

[4] B. Boeckmann, A. Bairoch, et al. "The SWISS-PROT protein sequence database and its supplement TrEMBL" *Nucleic Acids Res.,* 31(1), 2003, pp 365-370.

[5] T. Elomaa "In Defense of C4.5: Notes on Learning one-level Decision Trees" in 2003 *Proceedings of 11th Intl. Conf. Machine Learning*. Morgan Kaufmann. pp 62-69.

[6] R. Jensen, H. Gupta, et al. "Prediction of human protein function according to Gene Ontology Categories" in 2003 *proceedings of Bioinformatics,* 19, pp 635-642.

[7] L. Jensen, et al. "Prediction of Human Protein Function from Post-Translational Modifications and Localization Features" *Journal of Molecular Biology*, 319(5). 2002, pp 1257-65.

[8] L. Jensen "Prediction of Protein Function from Sequence Derived Protein Features" *Ph.D. thesis* 2002*, Technical University of Denmark*.

[9] D. Krane and M. Raymer, *Fundamental Concepts of Bioinformatics*, Benjamin Cumming: 2002.

[10] E. Kretschmann, W. Fleischmann and R. Apweiler "Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT" *Bioinformatics*, 17, 2001, pp 920-926.

[11] R. D. King, A. Karwath, A. Clare and L. Dehaspe "Accurate prediction of protein functional class in the M. tuberculosis and E. coli genomes using data mining" *Comparative and Functional Genomics*, 17, 2000 pp. 283-293.

[12] R. Kohavi and R. Quinlan "Decision Tree Discovery" *Data Mining*, 6, 2000,pp. 10-18.

[13] P. Adriaans and Zantinge, *Data Mining*, Pearson Education: 2002.

[14] D. Devos and A. Valencia "Practical Limits of Function Prediction. Protein Design Group", *National Centre for Biotechnology, CNB-CSIC,* 2000, Madrid, E-28049, Spain.

[15] J. Tamames et al. "EUCLID: Automatic classification of proteins in functional classes by their database annotations" *Bioinformatics*,1998, pp. 542-543.

[16] H. Almuallim, Y. Akiba, and S. Kaneda, "On handling tree-structured attributes in decision tree learning" in *Proceedings of the 12th International Conference on Machine Learning* (ICML95).

[17] J. Friedman, R. Kohavi and Y. Yun "Lazy decision trees" in *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, AAAI Press and the MIT Press,1994 pp. 717-724.

[18] J. R. Quinlan, "Induction of decision trees" *Machine Learning*, 1, 1993,pp 81-106.

[19] S. R. Safavian and D. Landgrebe "A Survey of Decision Tree Classifier Methodology" *IEEE Trans. Systems, Man and Cybernetics*, 21(3),1991, pp. 660-674.

Manpreet Singh received the B.Tech. Electronics & Electrical Communication from Guru Nanak Dev Engineering College, Ludhiana and M.Tech. in Computer Science & Engineering from P. A. U., Ludhiana. He is presently working with Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana. His current research interests are Bio-informatics, Distributed Computing and Data Mining. He has published around 20 research papers in various National and International conferences.

Parvinder Singh Sandhu is working as Assistant Professor in the Department of Computer Science and Engineering with Guru Nanak Dev Engineering College, Ludhiana (Punjab). He is Master of Engineering in Software Engineering, M.B.A. and Bachelor in Computer Engineering from National Institute of Technology (NIT), Kurukshetra. He has published 10 research papers in referred International journals and 15 papers in renowned international conferences. His current research interests are Software Reusability, Bio-informatics, Software Maintenance and Machine Learning.

Parminder Kaur Wadhwa received the B.Tech. in Computer Science & Engg. From Punjab Technical University and M.Tech. Computer Science & Engg with Gold-Medal from Guru Nanak Dev Engineering College, Ludhiana (2004-2006). He is presently working with Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana. His current research interests are Bio-informatics and Machine Learning.