Indian Journal of Science and Technology

Supplementary Article

Layered Approach for Predicting Protein Subcellular Localization in Yeast Microarray Data

A. Kumaravel[1*] and R. Pradeepa[2]

[1]Professor and Dean, Department of Computer Science and Engineering, Bharath University, Selaiyur, Chennai-600073, India; drkumaravel@gmail.com

[2]PG Student, Department of Computer Science and Engineering, Bharath University, Selaiyur, Chennai - 600073; pradimca@gmail.com

Abstract

Subcellular localization is a well-designed representation of proteins. We need a fully automatic and reliable prediction system for protein subcellular localization, especially for the analysis of large-scale of yeast microarray data. In this paper we consider the dataset with multi classes and propose the classification for each location of protein subcellular in a separate layer. In this work, a multi-classification approach for subcellular localization is designed and developed to achieve high efficiency and improve the prediction and classification accuracy. The rule based Ripper method has been found to predict the subcellular localization of proteins from their protein microarray data, compared to other classifiers.

Keywords: Data Mining, Microarray, Classification, Layered Approach, Protein Subcellular Localization.

1. Introduction

Genome function annotation including the assignment of a function for a potential gene in the raw sequence is now the hot topic in bioinformatics. Subcellular localization is a key functional characteristic of potential gene product such as proteins. Therefore, a fully automatic and reliable prediction system for protein subcellular localization would be very useful. Numerous stabs have been made to predict protein subcellular localization. Maximum of these prediction method scan be classified into two categories: one is based on the recognition of protein N-terminal sorting signals and the other is based on amino acid composition [1]. Recently, they proposed an integrated prediction system for subcellular localization using neural networks based on individual sorting signal predictions or Support Vector Mechanism for general purpose supervised pattern recognition. This paper introduces a new prediction method for protein subcellular localization based on yeast dataset. Classification is perhaps the most familiar and most popular data mining technique. Prediction can be thought of as classifying an attribute value into one of a set of possible classes. Here, we construct a prediction system for subcellular localization called Ripper based on the Data mining classification method. The results show that the prediction accuracy is significantly improved with this novel method and the method is very robust to errors in the yeast microarray data [2].

2. Problem Statement

Protein subcellular localization prediction is a multi-class classification problem. Here, the class number is equal to 10. A simple strategy to handle the multi-class classification is to reduce the multi-classification to a series of binary classifications. For a k-class classification, k Layers are constructed. The i^{th} layer will be trained with all of the samples in the i^{th} class with positive labels and all other samples with negative labels. We refer to this way as one-versus-rest. At Last one unknown sample is classified into class

* *Corresponding author:*
A. Kumaravel (drkumaravel@gmail.com)

that corresponds to the one-versus-rest with the highest output value.

2.1. The Proposed Layered-model Subcellular Localization

Our system has the capability of classifying micro array data. The data is input, which identifies if this record is in particular location or rest. If the record is identified as specified location then the module would raise a flag to the administrator that the coming record is a specified location which is taken into account and the other all locations are taken together, which consists of ten sequential Layers, one for each class type (ERL, NUC, POX, MIT, ME1, ME2, ME3, CYT, EXC, & VAC) [3]. Each Layer is responsible for identifying the location of coming record according to its class type. Else the locations are taken together as rest. Then each layer act as a filters that classifies the location of each layer category which eliminate the need of further processing at subsequent layers but we took in consideration the propagation of errors as to simulate the real system and results be more accurate and real [4] .

In many situations, there is a trade-off between efficiency and accuracy of the system and there can be various avenues to improve system performance. We implement the Layered Approach to improve overall system performance as our layered model using JRipRule achieves high efficiency and improves the detection and classification with high rate of accuracy. Figure 1 shows the proposed layered model of a subcellular localization.

3. Experimental Analysis and Results

In this section, first we collect the Microarray dataset. We apply the dataset in Wekatool [5] to find Classification results. The datasets for these experiments are from Institute of Molecular and Cellular Biology [6].

3.1 Dataset

3.1.1 Dataset Description

The database of Protein Localization is considered as our experimental dataset. It contains 10 different classes. Number of Instances in this database is 1484 for the yeast dataset and number of Attributes are 9 (8 predictive, 1 name).The class is the localization site shown in Figure 2.

3.1.2 Attribute Description

Attribute name	Description
Sequence Name	SWISS-PROT database accession number
mcg	Signal sequence recognition McGeoch's method.
gvh	Signal sequence recognition von Heijne's method.
alm	ALOM membrane spanning region prediction score.
mit	Discriminant analysis of the amino acid content of the N-terminal region score.
erl	Presence of "HDEL" substring binary attribute.
pox	Targeting Peroxisomal signal in the C-terminus.
vac	Discriminant analyses of the amino acid content of vacuolar and extracellular proteins score.
nuc	Discriminant analyses of nuclear localization signals of nuclear and non-nuclear proteins score.

3.2 Performance Evaluation

During the prediction of protein subcellular localization we observe two main challenging issues in this system. First, the number of locations in the subcellular is typically a very small fraction of the total microarray. Therefore the essential step is to select attributes of the various Layers. Second, the locations are classified in their impression and hence, it becomes necessary to treat them differently.

To improve the prediction rate, while maintaining a reasonable overall prediction rate. We have proposed a layered model with various classifiers [7] (BayesNet, NavieBayes, Decision Table and rules Jrip) on values. In layered model we define ten layers that correspond to the ten location groups i.e. NUC layer for predicting NUC location, MIT layer for predicting Probe location, CYT layer for predicting CYT location, EXC for EXC location and so on.

A confusion matrix illustrates the accuracy of the solution to a classification problem [8]. The accuracy and error rates are calculated from confusion matrix [9],

$$\text{Accuracy} = (TP+TN) / (TP+FN+FP+TN) *100$$
$$\text{Error rate} = (FN+FP) / (TP+FN+FP+TN) *100$$

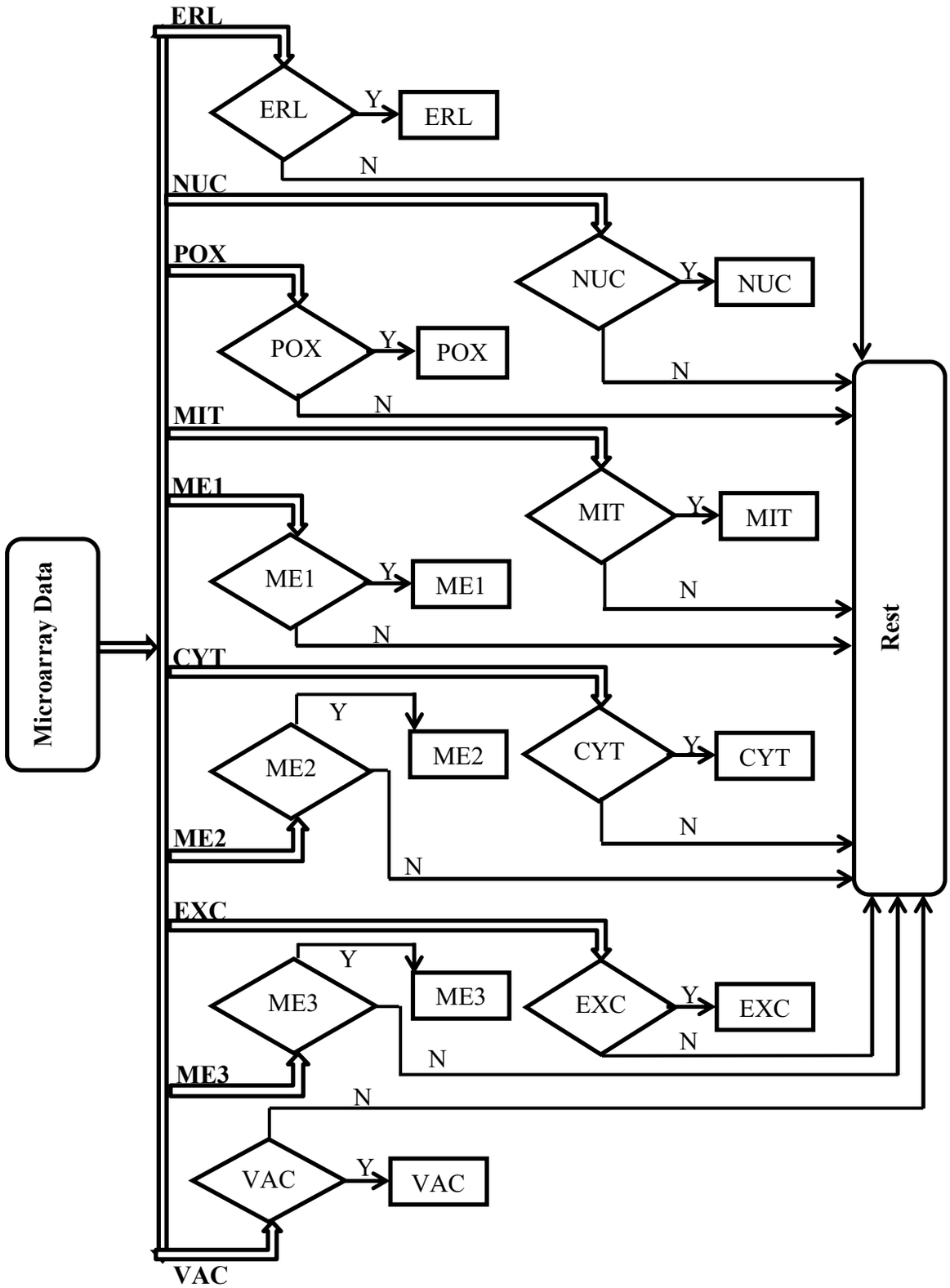

Figure 1. Layered-Model Approach System.

Column 1	Rest	EXC	Column 1	Rest	CYT
Rest	1441	8	Rest	1013	8
EXC	20	15	CYT	26	437
Error Rate		1.89%	Error Rate		27.90%

Column 1	Rest	MIT	Column 1	Rest	NUC
Rest	99	145	Rest	996	59
MIT	1161	79	NUC	36	393
Error Rate		12.00%	Error Rate		6.40%

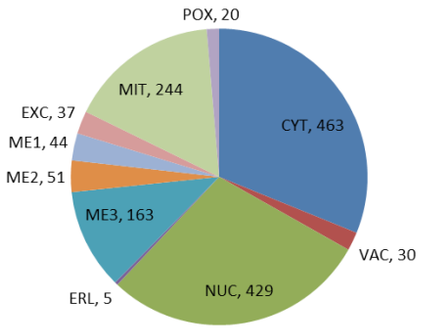

Figure 2. Location categories.

where, TP=True Positive, TN=True Negative, FP=False Positive, FN=False Negative

3.2.1 Classification of Layers

Records are classified [11] as locations by classifying the ten sequential layers; coming location to one of the ten classes (ERL, NUC, POX, MIT, ME1, ME2, ME3, CYT, EXC, & VAC) and identifying its type.

NUC Layer:

The results of NUC Layer are shown in Table 1.

MIT Layer:

The results of MIT Layer are shown in Table 2.

CYT Layer:

The results of CYT Layer are shown in Table 3.

Table 1. Classification of NUC layer

Method	Correctly classified	Incorrectly classified
bayes.BayesNet	93.60%	6.40%
bayes.NaiveBayes	88.07%	11.93%
rules.Jrip	79.31%	20.69%
rules.Decisiontable	77.76%	22.24%

Table 2. Classification of MIT layer

Method	Correctly classified	Incorrectly classified
bayes.BayesNet	87.00%	13.00%
bayes.NaiveBayes	88.00%	12.00%
rules.Jrip	87.00%	13.00%
rules.Decisiontable	86.65%	13.35%

Table 3. Classification of CYT Layer

Method	Correctly classified	Incorrectly classified
bayes.BayesNet	66.44%	33.56%
bayes.NaiveBayes	68.26%	31.74%
rules.Jrip	72.10%	27.90%
rules.Decisiontable	69.07%	30.93%

Table 4. Classification of EXC layer

Method	Correctly classified	Incorrectly classified
bayes.BayesNet	94.95%	5.05%
bayes.NaiveBayes	95.08%	4.92%
rules.Jrip	98.11%	1.89%
rules.Decisiontable	97.50%	2.50%

EXC Layer:

The results of EXC Layer are shown in Table 4.

The results of all Layers with best methods are shown in Table 5.

4. Performance Comparison with

Table 5. Layers with best classification method

Layer	Correctly classified	Incorrectly classified	Best method
MIT	88.00%	12.00%	NAVIBAYES
NUC	93.60%	6.40%	BAYES
CYT	72.10%	27.90%	JRIP
ME1	98.18%	1.89%	JRIP
ME2	96.56%	3.44%	JRIP
ME3	95.28%	4.72%	NAVIBAYES
EXC	98.11%	1.89%	JRIP
ERL	99.93%	0.07%	JRIP
POX	99.13%	0.87%	JRIP
VAC	97.98%	2.02%	JRIP

Table 6. Performance comparison with existing system

Methods	Neural network	Markov model	SVM	Proposed method
Location	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)
Cytoplasmic (CYT)	55	78.1	76.9	72.1
Extracellular (EXC)	75	62.2	80	98.1
Mitochondrial (MIT)	61	69.2	56.7	88
Nuclear (NUC)	72	74.1	87.4	93.6

Existing Approaches

In this section, we compare the performance of Ripper method predictions were compared with other prediction methods in this field[1]. This information is shown in Table 6.

According to the above table, the accuracy for Extra Cellular (EXC) sequences reached 98.1% with the rule based JRip method which is much higher than for the other methods.

5. Conclusions

A multi-Layer localization prediction system has been developed to achieve high efficiency and improve prediction and classification rate accuracy. The proposed system consists of two stages. First stage is for location detection and the second stage is for location classification. The data is input in the first Stage which identifies if this record is a specified record or rest.

Experimental results specify that the proposed layered model with JRip classifier can result in better prediction of minority classes without hurting the prediction performance of the majority class.

6. Acknowledgements

The authors would like to thank the management of Bharath University for the support and encouragement for this research work.

7. References

1. Hua S, and Sun Z (2001). Support vector machine approach for protein subcellular localization prediction, Bioinformatics, vol 17(8), 721–728, Available From: http://bioinformatics.oxfordjournals.org/
2. Available From: http://videocast.nih.gov/pdf/rm/Snyder.pdf
3. Gifty P, and Ravichandran J M et al. (2012). Efficient classifier for R2L and U2R attacks, International Journal of Computer Applications, vol 45, No. 21, 0975–8887.
4. Kumaravel A, and Niraisha M (2013). Multi-classification approach for detecting network attacks ICT545, 2013-IEEE Conference on Information and Communication Technologies.
5. Available From: Weka: http://www.cs.waikato.ac.nz/~ml/weka/
6. Protein localization dataset from Institute of Molecular and Cellular Biology Osaka, University, Available From: http://www.imcb.osaka-u.ac.jp/nakai/psort.html
7. Gaur A, and Richariya V (2011). A layered approach for intrusion detection using meta-modeling with classification techniques, International Journal of Computer Technology and Electronics Engineering (IJCTEE), vol 1(2), 161–168.
8. Kumaravel A, and Pradeepa R (2013). Efficient molecule reduction for drug design by intelligent search methods, International Journal of Pharama and Bio Sciences, vol 4(2): (B), 1023–1029, Available From: http://www.ijpbs.net/cms/php/upload/2348_pdf.pdf
9. Available From: http://webdocs.cs.ualberta.ca/~eisner/measures.html