GRADIENT DESCENT LOGIT BOOST CLASSIFICATION TECHNIQUE FOR DIAGONISING CANCER DISEASES ACCURATELY

I. Poonkody[1], A. Revathi [2] Dr. P. Sumathi [3]

Assistant Professor, Department of Computer Science,
New Prince Shri Bhavani Arts & Science College,
Medavakkam, Chennai. Tamil Nadu, India[1]

Assistant Professor, PG & Research Department of Computer Science and Applications,
Vivekanandha College of Arts and Sciences for Women (Autonomous),
Elayampalayam, Tiruchengode. Namakkal District. Tamil Nadu, India[2]

Assistant Professor, PG & Research Department of Computer Science,
Government Arts College,
Coimbatore-18. Tamil Nadu, India[3]

ABSTRACT

Biological data has been generated by the human genome and the sequencing projects for other organisms. The massive demand for analysis and interpretation of these data is being managed by the emerging science of bioinformatics. It is an interdisciplinary field, which connects computer science, mathematics, physics, biology and medicine. The data mining technique uses the protein data set to diagnose the various diseases. Due to less accuracy and more complexity, the need for better disease diagnosing technique arises. At this juncture, an improved disease diagnosing technique with less complexity and more accurate Ensembled Decision Tree with Gradient Descent Logit Boost Classification (EDT-GDLBC) technique was introduced. This paper describes the way to construct the decision trees which are base learners, that not only identifies the abnormal sequences based on their relationship between training and testing protein data sequences, but also diagnose diseases accurately with minimum time.

Keywords: Disease diagnosis, protein sequences, decision tree, bivariate correlation, gradient descent logit boost classifier

1. INTRODUCTION

Classification model receives greater attention in medical domain for disease diagnosis. Classification techniques are applied to various applications such as disease diagnosis, disease prediction, bioinformatics data analysis, and so on. In general, health care data are often huge and complex. Various data mining techniques have been developed for disease diagnosis using gene formation, DNA sequences, and protein sequences. But it failed to perform efficient prediction analysis of disease in certain cases. Therefore, an efficient classification is required for diagnosing the disease with protein sequences. An efficient Cost-Sensitive Classifier with Gentle Boost Ensemble (Can-CSC-GBE) technique was developed in [1] to identify the breast cancer using protein features. This technique does not minimize the computational complexity to identify the breast cancer using protein feature classification.

The Mega-Trend Diffusion (MTD) technique was developed in [2] with Support Vector Machine (SVM) and k Nearest Neighbor (KNN) for identifying the breast and colon cancers with protein sequences.

The MTD technique with SVM outperforms well when compared to KNN. But MTD - SVM model has a higher classification error since the classifier does not use any boosting technique to improve the accuracy.

A Random Forest tree induction algorithm was designed in [3] for detecting the three various lung cancer tumors.

A machine learning algorithms namely SVM, ANN and NB were introduced in [4] for predicting the types of lung tumor based on protein attributes.

The issues are identified from the above-said literature such as lack of diagnosing accuracy, more time and space complexity, incorrect identification of disease, and so on. In order to solve such kind of issues, an efficient ensemble classifier was introduced.

The major contribution of the EDT-GDLBC technique is described as follows,

- EDT-GDLBC technique provides a new contribution to diagnose the disease with minimum time. It constructs the number of base learners to classify the protein sequences as either normal or abnormal through the relationship
- The gradient descent logit boost classification technique is applied to improve the classification performance by combining the several base learners.

The remaining section of the article is organized in the following order. Section 2 provides an overview of related works. The proposed EDT-GDLBC technique is described in Section 3 with neat diagram and algorithm. In Section 4, Experimental evaluation of EDT-GDLBC technique and existing state-of-art methods are described with the dataset. Section 5 provides the results and discussion with certain parameters. Finally, Section 6 concludes the present work.

2. RELATED WORKS

Classification of protein sequences using the machine learning methods is a significant research in Bioinformatics. An ensemble fuzzy total margin support vector machine (EnFTM-SVM) classifier was developed in [5] for classifying the imbalanced protein data.

The issues identified from the above-said methods are rectified by introducing an ensemble classifier. The process of ensemble classification based disease diagnosis is presented in the next section

3. GRADIENT DESCENT LOGIT BOOST CLASSIFICATION USING ENSEMBLE DECISION TREE FOR DISEASE DIAGNOSIS

Disease diagnosis is the major concern in healthcare industry using protein sequences. Proteins sequences are constructed with the amino acids using peptide bonds. The mutation in the regular sequences causes diseases in a living organism. The identification of the disease is performed through the classification algorithms. But it has high prediction error while identifying the disease. In order to improve the diagnosing accuracy, Ean Efficient Ensemble Classifier was introduced. Ensemble classifier is a combination of the weak classifier to make a strong classifier which provides the accurate classification results rather than the single classifier. Based on the above motivation, Ensembled Decision Tree with Gradient Descent Logit Boost Classification (EDT-GDLBC) technique was introduced. In the proposed classification technique, decision trees are used as a weak classifier (i.e. base classifier) for diagnosing the disease with training samples. The gradient logit boost is a boosting technique which combines the entire weak learner and provides the stronger classification results with a minimum error. The architecture diagram of the proposed technique is introduced in figure 1.

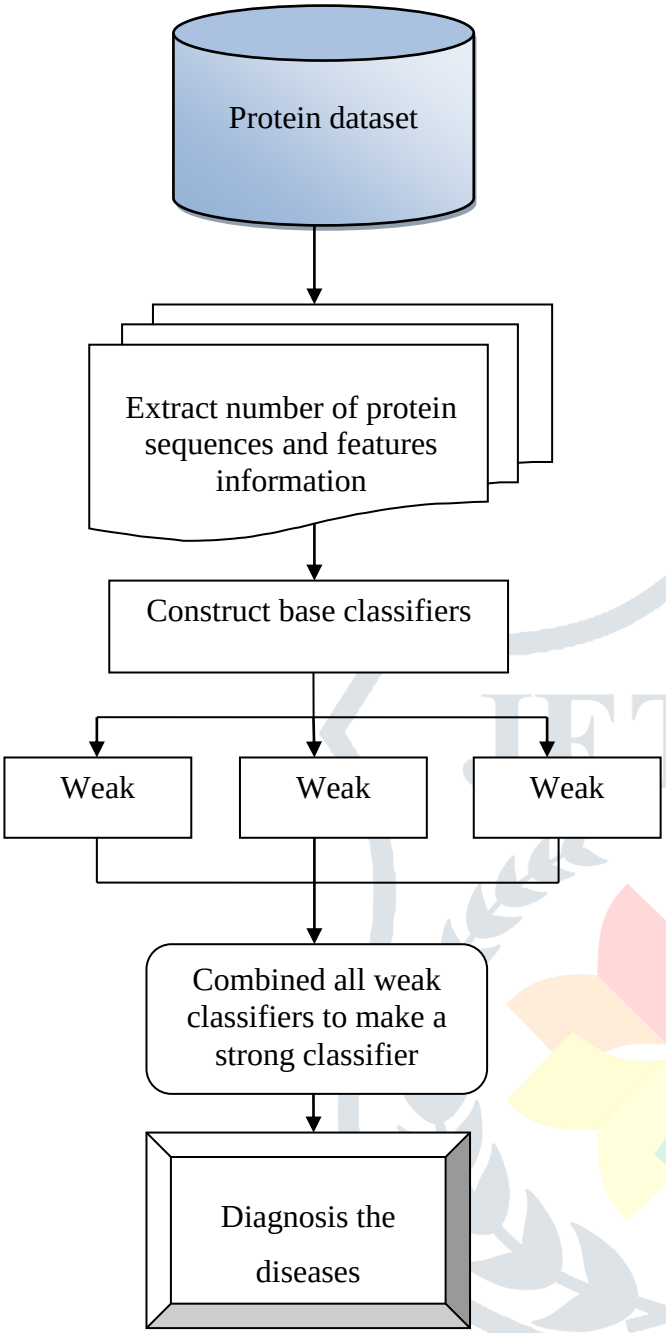

Figure 1 Architecture diagram of EDT-GDLBC technique

As shown in figure 1, architecture diagram of the EDT-GDLBC technique is described for diagnosing the disease. This is done by applying ensemble classifier. The number of protein sequences and their features information is extracted from the protein dataset. Based on their features, the normal and abnormal protein sequences are classified to identify the disease at an earlier stage. As a result of this classification, the dimensionality gets reduced as well as minimizes the wrong positive rate. The following section briefly discusses the ensemble classifier.

3.1 Ensembled Decision Tree with Gradient Descent Logit Boost Classification

The classification is the process where the similar data are categorized into different classes. The EDT-GDLBC technique performs the efficient classification by constructing the strong classifier. EDT-GDLBC technique constructs a number of base classifiers as a decision tree to classify the protein sequences. All these base classifier results are combined to provide the strong classification results using gradient descent logit boost classifier. Initially, the numbers of protein sequences in the dataset are expressed as follows,

$$P_{s_i} = p_{s_1}, p_{s_2}, p_{s_3}, \ldots . p_{s_4} \in D^p \text{ ----------(1)}$$

From (1), p_{s_i} denotes a protein sequences, D^p denotes a protein dataset. Each sequences has different features and it expressed as follows,

$$P_{s_i} \rightarrow f_1, f_2, f_3, \ldots f_n \in D^p \text{---------- (2)}$$

From (2), $f_1, f_2, f_3, \ldots f_n$ denotes features in the dataset. With the given input protein sequence and their feature information, the decision tree is constructed for classification. In the proposed technique, a Decision tree builds classification models in the form of a tree structure. It divides a dataset into smaller subsets. The main advantage of the decision tree classifier is to handle both categorical and numerical data.

4. RESULTS AND DISCUSSION

In this section, the experimental results and discussion of EDT-GDLBC technique and two existing methods Can-CSC-GBE [1] and MTD-SVM model [2] are described with various performance metrics such as disease diagnosing accuracy, computation time, wrong positive rate and the space complexity. With the help of these parameters, the experimental results are compared with the table and graphical representations.

4.1 Performance results of disease diagnosing accuracy

Disease diagnosing accuracy is defined as the ratio of the number of (i.e. no. of) protein sequences classified as active or inactive to the total number of protein sequences in the protein dataset. The

mathematical formula for disease diagnosing accuracy is expressed as follows,

$$DDA = \frac{No.of\ protein\ sequences\ \ are\ correctly\ classified}{n} * 100 \quad (11)$$

From (6), DDA denotes a Disease diagnosing accuracy and 'n' denotes a number of protein sequences taken as input for experimental evaluation. DDA is measured in terms of percentage (%).

Sample mathematical calculation for disease diagnosing accuracy

EDT-GDLBC: Number of protein sequences correctly classified is 971 and the number of protein sequences is 1000. Then,

$$DDA = \frac{971}{1000} * 100 = 97\%$$

Can-CSC-GBE: Number of protein sequences classified is 900 and the number of protein sequences is 1000. Then,

$$DDA = \frac{900}{1000} * 100 = 90\%$$

MTD-SVM model: Number of protein sequences classified is 725 and the number of protein sequences is 1000. Then,

$$DDA = \frac{725}{1000} * 100 = 73\%$$

Table 1 Tabulation for Disease diagnosing accuracy

No. of protein sequences	Disease diagnosing accuracy (%)		
	EDT-GDLBC	Can-CSC-GBE	SCMTD-SVM Model
1000	97	90	73
2000	86	78	66
3000	85	75	67
4000	81	75	69
5000	92	87	71
6000	94	83	66
7000	90	73	67
8000	89	82	74
9000	93	88	78
10000	96	90	81

Table 1 clearly describes the performance results of disease diagnosing accuracy versus number of protein sequences in the P53 mutant dataset [1] and MTD-SVM model [2]. The above results clearly illustrates an accuracy of proposed EDT-GDLBC technique is higher than the existing methods. This significant improvement is achieved by performing the ensemble classification. The EDT-GDLBC technique considers input protein sequences and their features information's from the dataset. Then the base classifier i.e. decision trees are constructed to classify the inactive protein sequences i.e. cancerous sequences.. Then the base classifier takes a decision i.e. if the training sequences information is matched with the testing results. If it is positive, then the normal sequences are identified otherwise it is said to be an inactive sequences. Finally, the logit boost ensemble classification is applied to combine all the base classifiers. By this way, the EDT-GDLBC technique identify and diagnosis the cancer disease with abnormal sequences. An experimental result shows that the disease diagnosing accuracy is increased by 10% and 27% when compared to existing Can-CSC-GBE [1] and MTD-SVM model [2] respectively.

5. CONCLUSION

An efficient technique called Ensembled Decision Tree with Gradient Descent Logit Boost Classification (EDT-GDLBC) is developed for diagnosing the disease at an earlier stage with higher accuracy. Based on the relationship, the decision tree classifies the protein sequences as normal or abnormal. All the weak decision trees are combined into strong classifier by using gradient descent logit boost classification for diagnosing the disease with minimum error. Experimental evaluation of EDT-GDLBC technique and state of the art methods are performed using P53 mutant dataset. Finally, the technique improves disease diagnosing accuracy with minimum time.

REFERENCES

[1] Safdar Ali,Abdul Majid , Syed Gibran Javed, Mohsin Sattar, "Can-CSC-GBE: Developing Cost-sensitive Classifier with Gentleboost Ensemble for breast cancer classification using protein amino acids and imbalanced data", Computers in Biology and Medicine, Elsevier, Volume 73, 2016, Pages 38–46

[2] Abdul Majid, Safdar Ali, Mubashar Iqbal, Nabeela Kausar, "Prediction of human breast and colon cancer from imbalanced data using nearest neighbor and support vector machines", Computer Methods and Programs in Biomedicine, Elsevier, Volume 113, Issue 3, March 2014, Pages 792-808

[3] Faezeh Hosseinzadeh, Mansour Ebrahimi , Bahram Goliaei, Narges Shamabadi, "Classification of Lung Cancer Tumors Based on Structural and Physicochemical Properties of Proteins by Bioinformatics Models", PLoS One, Volume 7, Issue 7, Pages 1-8

[4] Faezeh Hosseinzadeh, Amir Hossein KayvanJoo, Mansuor Ebrahimi and Bahram Goliaei, "Prediction of lung tumor types based on protein attributes by machine learning algorithms", Springer Plus, Volume 2, Issue 238, 2013, Pages 1-14

[5] Hong-Liang Dai, "Imbalanced Protein Data Classification Using Ensemble FTM-SVM', IEEE Transactions on Nano Bioscience, Volume 14, Issue 4, 2015, Pages 350 – 359

[6] Hong-Liang Dai, "Imbalanced Protein Data Classification Using Ensemble FTM-SVM', IEEE Transactions on Nano Bioscience, Volume 14, Issue 4, 2015, Pages 350 – 359