THE American Journal of Humanities and Social Sciences Research (THE AJHSSR) **2021**

E-ISSN: 2581-8868
Volume-04, Issue-01, pp-15-21
www.theajhssr.com
Research Paper Open Access

Improved AdaBoost Algorithm for Big Data Analysis: A Review

Susanta Kumar Sahoo[1], Sasmita Mishra[2], Dillip Kumar Swain[3]
[1,2,3]*Department of CSE&A, IGIT Sarang, Odisha, India*

ABSTRACT

AdaBoost is a popular machine learning technique which is represented as a boosting algorithm that introduced in 1995 by freund and schapire. It is handy in many machine learning application where a classifier with high level of accuracy can be applied and interpreted to many factors or rules to produce optimized result or can be applied for predicting the output. AdaBoost is a popular classifier which combines many weak learners to obtain and present a strong classifier. The weak learner may be SVM, decision tree, decision stump, logistic regression etc. AdaBoost provides a framework to combine weak learner (which should be able to handle weighted training examples) to obtain final classifier whose accuracy is significantly higher than accuracy of many single model. Many time imbalanced data exhibit an unequal distribution among its categories. In this case classification rule plays a major role classifying these data sets. Weak classification rules are always a problem. So in this case improved AdaBoost may incorporate many classification rules to an improved and upgraded classifier to classify the data set. In this paper we have provided a review upon the application of AdaBoost upon big data analysis. We have also discussed its pros and cons upon combining many weak learning techniques and the detail working principle on combining many learning techniques. Similarly protein structure prediction is one of the important tasks of bioinformatics which is the prediction of secondary, tertiary and quaternary changes that occur to the primary protein structure form time to time. As the genomic data folded many times in the last few decades so classifying different data set is a major task that should be processed with proper classification rules with appropriate learning technique. We have also applied AdaBoost for protein secondary structure prediction with combining Fuzzy support vector machine.

KEYWORDS : AdaBoost, Learning Algorithm, Fuzzy support vector machine, Bioinformatics, Protein, Soft computing

I. INTRODUCTION

Classification is the basic idea behind every data mining technique to collect the important research contents. Generally almost all the classifiers based upon some hypothesis. As the amount of data sets and information are folded many times in the last few decades so a proper classification rule is very important to design in order to get the optimized result. However it's always a challenging task to design a suitable classifier for a big data analysis i.e. when the data sets are very big. So it is wise to use more than one classification rules or more than one classifier to execute the task. Again using more and more classifier creates a complex platform during the data manipulation and result abstraction. Improved AdaBoost is such a popular technique which provides a good platform and framework to combine more classification methodology to obtain meaningful results in a data mining process. In data mining, some of the classification methods are relatively mature. These classification rules perform well in classifying balanced data. Imbalanced data exists in many fields like information retrieval, fraudulent activities on credit cards, medical diagnosis etc. minority class is more important than majority class in these areas. For an example we can consider a case from medical science point of view. If a healthy person is misdiagnosed as patient in medical diagnosis, it will bring him a burden and a bad influence on doctor; If a patient is diagnosed as healthy person, then it may miss the best treatment period, which will cause critical illness [2]. In boosting system many week learners are consolidated and gives strength to build up a new improved learning technique. Strong learners are those classifiers which give maximum accuracy and may considered as the base of machine learning. Strong learners have wide applications in vast areas and it can be applies on many classification techniques like feature extraction, feature selection, multi-class categorization etc.

This paper is organised as follows. The chapter I is the Introduction which provides an overall discussion about AdaBoost, Protein Structure prediction, big data analysis and its applications. Chapter II is the literature review which provides a brief idea and discussion about the current and past research and developments of AdaBoost along with other machine learning algorithms. The Chapter III provides an idea about big data analytics and its necessity in this day to day life. The Chapter IV provides the detail working principle, algorithms and methodology about application of AdaBoost along with FSVM and Hyper plane optimization [5] for PSP problems. The Chapter V provides the Simulation results and a salient discussion about the findings. Finally the paper concludes in chapter VI with conclusion and future work.

II. ADABOOST AND ITS APPLICATIONS

AdaBoost has several applications for data classification, analysis and result publication purpose. Adaptive learning is shortly abbreviated as AdaBoost which is the most commonly used machine learning algorithm. The basic technology behind boosting system is various weak learners are consolidated and give a strong learner with higher precision. These weak learners are those whose prediction capacity is more accurate than that of random guessing. Strong learners are the classifiers which give maximum accuracy and there may not be necessity of integrating with other classifying technique. AdaBoost has wide application area and can be applied on many classification techniques i.e. feature selection, feature extraction, and multi-class categorization etc. Some of the most widely used applications of boosting include text classification, medical application, academic and commercial etc.

Ada Boost Applications in Data Analysis : Boosting technique introduced by Freund and Schapire is a type of ensemble technique which is used with a collection of many weighted same or different type of predictors. However in other way several hypotheses are selected and eventually their prediction result is combined. For example, if 50 decision trees are generated over same or different training data set then a new test dataset is created and voted for best classification. So in this base learner is chosen and improved it iteratively for the misclassified data [4]. In short AdaBoost is:

- Assign equal weight to all training data.
- A base algorithm is chosen.
- At each step, increase the weight of misclassified data.
- Iterate it n times.
- The Final model is made by weighted sum of n learners [4].

XGBoost one more application of AdaBoost stands for extreme gradient boosting. This may considered as an implementation over the gradient boosting. Greedy approach is applied for XGBoost which has a good performance and speed. It is also having the following advantages.

- Missing Values: XGBoost has built-in function that handles missing values.
- Speed: Due to parallel processing process it has faster performance than gradient boosting.
- Remove over fitting: It controls the over fitting problem [4].

PSP problems and its necessity : Proteins are the large biological molecule in a living body which is a combination of twenty different amino acids with variation in their percentages. Protein has the various works in a living body like catalyzing metabolic reaction, responding to stimuli etc. Protein structure prediction is the prediction of the three dimensional structures that is commonly framed by the amino acid sequence. The three dimensional structure framed by the amino acid compositions and it generally changes its shape due to the effect of external agents or drugs to these amino acid compositions. The necessity of prediction of protein structure is to design new drugs or medicines for the cure and treatment of patients. From the protein structures the medicine researchers and doctors working for the design of medicine and drugs can easily examine the improvement, changes that occur during the treatment of a patient after application of medicines [3].

The amino acid composition generally creates different shapes in its three dimensional shapes. The shapes may be in the form of α-helix, β-sheets or loop etc. Similarly twin structure removal is also another important task during protein structure prediction. It is very much important for removal of twins as well as the matching structures. Twin removal [3] is also provides a very good solution to the problems of removing similar individuals from the population. But as these types of data sets are very large in size and in the last decades these are folded into many times so proper classification technique is required for the removal of twin structures. Applying AdaBoost with number of regular classification technique for these problems also provides more optimized result than that of a simple classification technique [3].

III. BIG DATA ANALYTICS: A REVOLUTION OF MODERN TIME

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with large size and complexity that traditional data management tools may not store it or process it efficiently. So it is always a challenge for various researchers to process these types of data in an efficient manner with proper classification technique.Various computational intelligence methodologies are playing a major role in Data Analytics including gaining knowledge from unstructured sequence data. Capturing and analyzing valuable information's from Computational Biology and classifying the structure of the protein, especially secondary structures from its sequence is crucial as the structure in turn identifies the function, which is considered as an important task in bioinformatics. In the last decades the feature extraction becomes a critical task and not suitable for this protein structure prediction problem. In [20] a deep learning technique has been proposed that is implemented in Distributed Framework for improved accuracy and performance. Skip-Gram method is used to translate the amino acid sequence into words, without losing the position information of each amino acid. In the next step the vector is then fed into Stacked Auto Encoder for classification and the classifier output predicts the presence of secondary structures [20]. Though Proteins are very large in size and these type of data are folded into many types so Big data analysis methodology plays a vital role in performing analysis for mining in psp problems.

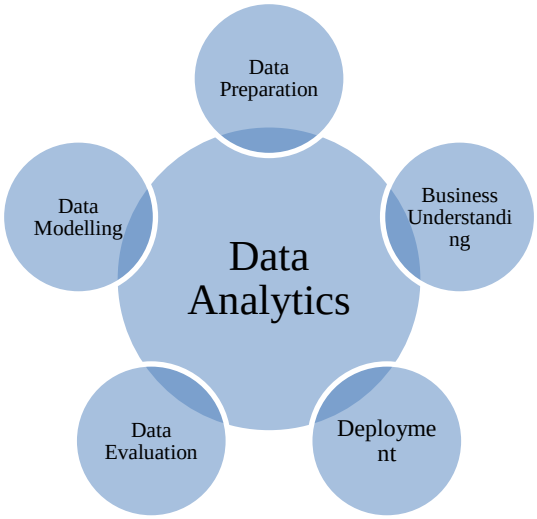

Fig. 1 Analysis wing of Data Analytics

Big data analysis is an essential science of the modern days. Starting from business modeling to the space science research data analysis and result analysis is essential. Big data analytics has several advantages like data preparation, Business understanding, Data modeling, data Evaluation, Deployment etc. As shown in Fig. 1 Data modeling is an important advantage of big data analytics. Collecting the meaningful and essential data and information from big collection is the intension before modeling any data model.For business modeling data analytics and business prediction is an important feature. Though lot of research are being carried out in this field but every day taking the size of data in to consideration these data need s proper prediction and classification technique for the optimized result. Researchers in the field of data mining always try to find innovative techniques so as to improve the performance of the extraction methods used in data mining as they usually use history of the different transactions done in finding the data as it will be useful for future use. AdaBoost is a powerful algorithm for classification that is widely used other fields like biology, speech processing etc. Instead of using the soft computing techniques like SVM and Fuzzy Sets, AdaBoost can achieve similar classification results with less effort and modifications in the parameters [8].

Classification is the technique which uses data to generate a model and assigns data items to one of the several distinct categories. This machine leaning method is capable of processing large amount of data. But many times when data sets are very big a simple yet a single classification may not sufficient to predict or to generate a business model. In that case AdaBoost is an alternative to modify the classification procedure by adding or by merging several classifiers or several week classifiers to generate a stronger one. In such cases the AdaBoost may be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data set also. These models may be used to make predictions based on the data items which can be used in businesses like hospitals, banks or any e-commerce business application also. In [8] V.K.P.German et al.

proposed the AdaBoost algorithm for the sales classification and prediction. The algorithm has capability to process both variable and numerical values. The first part of the algorithm converts the facts and represents them in numerical, computable values. Resulting to this the next part of the algorithm process the entire set of data numerically, which results in performing better for the algorithm. The basic intension of the technique is to improve the performance of the extraction methods used in data mining which is important for any type of business modeling or decision is making [8].

IV. APPLICATION OF ADABOOST FOR PSP PROBLEMS

The deep knowledge of the structural class of a given protein is important for understanding its folding patterns. A lot of efforts have been made; it still remains a challenging problem for prediction of protein structural class solely from protein sequences. The feature extraction and classification of proteins are the main problems in prediction. In protein feature extraction, AdaBoost technique can be applied by word frequency and word position collections from sequences of amino acid, reduced amino acid, and secondary structure. For accurate classification of the structural class of protein, a novel Multi-Agent Ada-Boost (MA-Ada) is being proposed method by integrating the features of Multi-Agent system into Ada-Boost algorithm. Extensive experiments were taken to test and compare the proposed method using four benchmark datasets in low homology. The results showed classification accuracies of 88.5%, 96.0%, 88.4%, and 85.5%, respectively, which are much better compared with the existing methodology [19].The data set are very big in size in case of biological data for its diversity. So processing of big size data set is very difficult during the experiment. For experiment purpose we have taken the small database 4HHB. During the simulation the amino acid composition of a total of 2 proteins is shown below. The pictorial representation of this protein is shown in Fig.2. Considering the percentage of all amino acid composition, Leucine (L) and Isoleucine(I) is having the Highest and lowest amino acid composition respectively. The amino acid compositions are as follows:

Number of proteins in database: 2

Amino acid composition in database:

F	5.2265%	S	5.5749%
T	5.5749%	N	3.4843%
K	7.6655%	E	4.1812%
Y	2.0906%	V	10.8014%
Q	1.3937%	M	1.0453%
C	1.0453%	L	12.5436%
A	12.5436%	W	1.0453%
P	4.8780%	H	6.6202%
D	5.2265%	I	0.0000%
R	2.0906%	G	6.9686%

Enzyme cleaved fragments:

Non redundant fragments 69

Redundant fragments 0

Out of range fragments 12

Total time: 0 sec

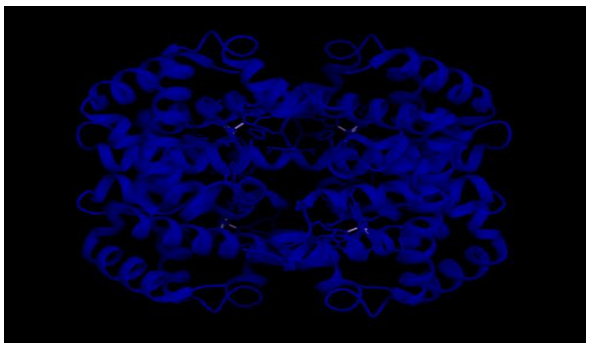

Fig. 2 Protein Structure of 4HHB.

According to the science of living things proteins are the large molecules available in a living body. As Protein data sets are very large in size and processing take more times so AdaBoost may be applied for the prediction and processing of the Protein Molecule. Here we have applied AdaBoost for the data set 4HHB for classifying the α-helices, β-sheet and loop content in the molecule. AdaBoost along with Fuzzy Support vector machine classification technique has been applied for differentiating the type of structure and percentage of structure present in the protein molecule. When using FSVM for the protein secondary structure prediction, the construction of an appropriate membership function is the basic key point. Generally, the fuzzy membership function is set according to the distance between the sample point and its class center in the input space. SVM has wide applications in bioinformatics and other related fields. Generally SVM deals equally with the training samples when applied to predict the secondary structure, and sometimes may also cause the over-fitting problem. During the use of FSVM for the protein secondary structure prediction, the construction of appropriate membership function is the major requirement. Normally the fuzzy membership function is set according to the distance between the sample point and its class center. Before the classification of data, SVM needs to map the raw input data into a high-dimensional feature space [5]. Therefore, the performance of FSVM based on these kinds of fuzzy memberships will be unsatisfactory. To address this problem, a new membership function calculated in the feature space has been proposed and defined as

$$S_1 = 1 - \sqrt{d_1^2/(\,r_p^2 + \delta}$$

where r_p is the radius of the class Cp defined as $\max_{X_i \in C_P} || \phi_P - \phi(X_i)||$, d_i^2 is the square of the distance between the trainin sample $X_i \in C_P$ and its class centre which is defined as $d_i^2 = K(X_i, X_i) - \frac{2}{n_p}\sum_{X_j \in C_p} K(X_i, X_j) + \sum_{X_j \in C_p} \sum_{X_k \in C_p} K(X_j, X_k)$ the corresponding class centre of class C_p is defined as $\phi_P = \frac{1}{n_p}\sum_{X_i \in C_p} \phi(\,X_i)$ where n_P is the number of samples in class C_p, $K(X_iX_J) = \phi(X_i)^T\phi(X_{J)}$ is a kernel function and δ is a small number to avoid the case when S_i is equal to 0. All these expressions are calculated in the feature space [5]

V. SIMULATION AND RESULTS

Protein Structure Prediction is one of the important tasks of Bioinformatics. Though Several research and results are in this filed but time to time due to the availability of new biological data and methodology this field attracts may researcher to work upon this field i.e. in bioinformatics and each day new technique are being designed. In our experiment we have taken the data set 4HHB. We have used Fuzzy Support vector machine along with hyper plane optimization for the classification of different structures. The 4HHB is having two protein molecules. The sequence of the twenty amino acid composition of the two protein is as follows:

4HHB_1|ChainsA,C|Hemoglobin subunit alpha|Homo sapiens (9606)

VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADAL
TNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVST
VLTSKYR

4HHB_2|ChainsB,D|Hemoglobin subunit beta|Homo sapiens (9606)

VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKK
VLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKV
VAGVANALAHKYH

Similarly we have performed the classification regarding the percentages of different structures available in this data set. The Fig.3 represents the content of 4HHB. The α-helix content is little bit more than that of β-sheet and others. It contains the α-helix percentage of 58% and β-sheet percentages of 25% and other structures having 17%. AdaBoost is very useful for combining several week classifiers for the journey towards an optimized result. The 4HHB belongs to homo sapiens used the method X-RAY diffraction with resolution 1.74 Å [21]. The AdaBoost can be used for large data sets for other bioinformatics applications like DNA-RNA alignment and comparison, Gene Mapping and other data processing works. It works best for big data analysis and processing with more accurate result than that of application of any single classifying technique. .

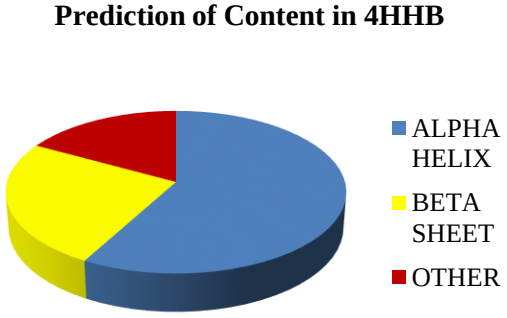

Fig. 3 Prediction for 4HHB

Research Collaboratory for Structural Bioinformatics (RCSB) is one of the best organizations which provide these types of molecular data sets for the research purpose. Several data sets that are collected with different experiment are available in free of cost. 4HHB is a small yet very paramount data set for protein structure prediction point of view. The AdaBoost is best suited for large data sets in this scenario. It will work efficiently for the protein structure prediction yet processing of large protein data set for the calculation of amino acid composition

VI. CONCLUSION AND FUTURE WORK

Though Applying AdaBoost is little bit old application but it has several effects with the new and improved classifier. Now days as the data set are very big in size so it is always the challenge to collect the meaningful data and mining information's. So in this aspect the AdaBoost is very useful with combining several classifying technique to improve the mining quality with improving the traditional classifier. In this paper our approach of applying AdaBoost for protein structure prediction will play a vital role for protein structure prediction. Applying AdaBoost with combining Fuzzy support vector machine with a hyper plane optimization provides better prediction with optimized result when the comparison is in between more than one protein molecules of same species. This technique will be helpful for the prediction of protein structure along with calculation of α-helics, β-sheet and loop percentages. Our upcoming research will focus upon the DNA-RNA alignment and comparison using AdaBoost. Applying AdaBoost i.e. applying Fuzzy support vector machine along with other classifying techniques will be helpful for various DNA-RNA alignment and classification of various genomic transformations.

REFERENCES

[1] K. Li, P. Xie, J. Zhai, W. Liu, "An Improved Adaboost Algorithm for Imbalanced Data Based on Weighted KNN" Proc. of 2nd Int. Conf. on Big Data Analysis, IEEE, pp. 30-34, 2017.

[2] G.M. Weiss, F. Provost, "Learning when training data are costly: the effect of class distribution on tree induction" Journal of Artificial Intelligence Research, Vol. 19, pp.315-354, 2003.

[3] S. B. Rout, S. Mishra, S.N. Mishra, "Protein Structure Prediction for Dialysis Patients using Genetic Algorithm" International Journal of Scientific Research in Computer Science Applications and Management Studies, Vol.8(1), 2019.

[4] A. N. Sharma, "Survey of Boosting Algorithms for Big Data Applications" International Journal of Engineering Research & Technology, Special Issue, 2017.

[5] S. Xiea, Z. Lia, H. Hua, Protein secondary structure prediction based on the fuzzy support vector machine with the hyperplane optimization , Journal of Gene, Vol. 642 pp. 74–83, 2018.

[6] V. Chang, T. Li, Z. Zeng, "Towards an improved Adaboost algorithmic method for computational financial analysis", Journal of Parallel and Distributed Computing, Vol. 134, pp. 219–232, 2019.

[7] S. B. Rout, S. Mishra D. K. Swain, "Protein Structure Prediction of Amino Acid Compositions using Genetic Algorithm", International Journal of All Research Education and Scientific Methods, Vol.8(12), 2020.

[8] V. K. P. German, B. D. Gerardo, R. P. Medina, "Implementing Enhanced AdaBoost Algorithm for Sales Classification and Prediction" Int. Journal of Trade, Economics and Finance, Vol. 8(6), 2017.

[9] D. K. Swain, S. N. Mishra, S.B.Rout, "Privacy Preservation in Distributed Data Mining for Protein Secondary Structure Prediction" Proceedings of 2nd International Conference on Communication and Electronics Systems, IEEE, 2017.

[10] S. Wu, H. Nagahashi, "Analysis of Generalization Ability for Different AdaBoost Variants Based on Classification and Regression Trees" Journal of Electrical and Computer Engineering, Vol.38, 2015.

[11] S. B. Rout, S. Mishra, S. N. Mishra, "A Review on Application of Artificial Neural Network(ANN) on Protein Secondary Structure Prediction", Proceedings of Second IEEE Int. Conf. on Electrical, Computer and Communication Technologies, IEEE, 2017.

[12] M. Joshi, V. Kumar, R. Agarwal, "Evaluating boosting algorithms to classify rare classes: Comparison and improvements", Proceedings of the 1st IEEE Int. Conf. on Data Mining, pp.257-264, 2001.

[13] P. Viola, M. Jones, "Fast and robust classification using asymmetric AdaBoost and a detector cascade", Advances in Neural Information Processing Systems" pp.1311-1318, 2002.

[16] G. Karakoulas G, Shawe, J. Taylor "Optimizing classifiers for imbalanced training sets", Proceedings of the 1998 Conf. on Advances in Neural Information Processing Systems, pp. 253-259, 1999.

[17] N. V. Chawla, A. Lazarevic, L. O. Hall, "SMOTEBoost: improving prediction of the minority class in boosting", Proceedings of the 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases, pp. 107-109, 2003.

[19] M. Fan, B. Zheng, L. Li, "A novel Multi-Agent AdaBoost algorithm for predicting protein structural class with the information of protein secondary structure" Journal of Bioinformatics and Computational Biology, Vol. 13(6), 2015.

[20] X. Leo Dencelin, T. Ramkumar, "An approach to predict protein secondary structure using Deep Learning in Spark based Big Data computing framework, 4th Int. Conf. on Applied and Theoretical Computing and Communication Technology, 2018.

[21] G Fermi, M F Perutz, B Shaanan, R Fourme, "The crystal structure of human deoxyhaemoglobin at 1.74 A resolution" Journal of Molecular Biology, Vol. 175(2), 1984.

[22] S. B. Rout, S. Mishra, D. K. Swain, "Application of Genetic Algorithm in Various Bioinformatics Problems", International Journal of Innovative Research in Technology, Vol. 4(9), 2018.

[23] S. K. Sahoo, S. Mishra, D. K. Swain, "An Analysis of Performance of Multidimensional Stock Exchange Data using k-means Clustering" Think India Journal, Vol. 22(14), 2019.