International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249 – 8958 (Online), Volume-9 Issue-5, June 2020

Classification of Hot Spots using XGBoost and LightGBM Algorithms

Minul Vijayakumar, Joby George

Abstract*: *Protein-Protein Interactions referred as PPIs perform significant role in biological functions like cell metabolism, immune response, signal transduction etc. Hot spots are small fractions of residues in interfaces and provide substantial binding energy in PPIs. Therefore, identification of hot spots is important to discover and analyze molecular medicines and diseases. The current strategy, alanine scanning isn't pertinent to enormous scope applications since the technique is very costly and tedious. The existing computational methods are poor in classification performance as well as accuracy in prediction. They are concerned with the topological structure and gene expression of hub proteins. The proposed system focuses on hot spots of hub proteins by eliminating redundant as well as highly correlated features using Pearson Correlation Coefficient and Support Vector Machine based feature elimination. Extreme Gradient boosting and LightGBM algorithms are used to ensemble a set of weak classifiers to form a strong classifier. The proposed system shows better accuracy than the existing computational methods. The model can also be used to predict accurate molecular inhibitors for specific PPIs.

Keywords: Extreme Gradient Boosting (XGBoost), Protein Protein Interaction (PPI), Protein Protein Interaction Network (PPIN, LightGBM.

I. INTRODUCTION

Proteins are large biomolecules which are made of one or more long amino acid chains. Proteins normally associate with other proteins to form larger complexes for conducting biological functions [1,2]. They execute different tasks in a given time and place, which are the base of life activities [3,4]. Protein molecules have several types of interactions, including bonding-hydrogen interaction, ion-ion interaction [5]. Regardless of the effect of these connections the protein molecules fold from their primary structures to shape the 3D structures [6]. Proteins bind through the folding mechanisms to certain molecules, and they associate exclusively at different active sites with other molecules. In protein protein interactions (PPIs), the target molecules are typically certain forms of molecules, such as nucleic acid [7]. The specific protein associations are defined by the functional amino acid groups within the active sites [8].Discovering protein interaction with other proteins or DNA molecules improves our knowledge of cell biology on a broad scale and biological pathways. Understanding of protein-protein interactions is also important for learning more about the function of the protein structure, which remains a difficult activity in bioinformatics, bioscience and computer science. Interactions between proteins have a significant effect on virtually all biological processes. The aberrant experiences form the cause of lethal diseases, such as Alzheimer's disease (AD), Creutzfeldt-Jakob disease (CJD), or even cancer. There are several essential residues called hot spots which are a small fraction of residues in interfaces and contribute the significant binding energy in protein interactions. To understand the biological underpinnings of the disease, it is also crucial to recognize critical residues involved in protein-protein interactions. Indeed, the identification of hot spots or hot areas provides a possible foundation for the detection and study of molecular pharmaceutical products and diseases. A vast volume of evidence is developing on protein-protein interactions with the creation of high-throughput technologies. Several PPIN models have been established to better explain the biological behavior of cells and the complex shifts in biological processes [9]. Such PPIN structures are applied focused on a detailed knowledge of individual residues in amino acids. These residues of amino acids play a key role in the interactions between protein and protein. A conventional innovation in molecular science has been utilized to distinguish problem areas, known as alanine scanning, which decides the commitment of vital buildups to the security or the capacity of a given protein. The energy contribution of the system amino acids is calculated by mutating each amino acid to alanine. Yet this approach does not extend to large-scale deployment because the treatment is particularly costly and time intensive. The effective computational methods to classify hot spots are therefore greatly needed.An increasing number of researchers have suggested various statistical approaches to detect hot spots and hot regions in the protein and protein interactions. Kortemme developed a concrete model to precisely identify the hot spots. Ofran et al implemented the ISIS system of finding hot spots dependent on functionality. We also used a popular system prediction approach to classify hot spots dependent on amino acid sequences.

Revised Manuscript Received on June 15, 2020.
* Correspondence Author
Minul Vijayakumar*, Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. Email: minultv@gmail.com
Joby George, Associate Professor and Head of the Department of Computer Science and Engineering of Mar Athanasius College of Engineering, Kothamangalam, Kerala, India.

Darnell et al designed the prediction models to classify hot spots, based on historical awareness of the general existence of PPIs. They also implemented the KFC Database to imagine the dynamic world surrounding hot spots clearly.

While such analytical approaches have become extremely useful, certain problems also need to be addressed. Training methods, for example, have trouble achieving good classification efficiency for various samples. Specialists also saw that unique PPIN proteins, known as hub proteins, could be firmly related. At present, the best possible comprehension of the topological structure of hub protein is as yet a major test in PPIs. The hot spots would be linked to specific protein partners in various hub protein interfaces. Predicting the hot spots of hub protein interfaces is very useful for finding a protein that can bind to different protein partners. This paper therefore explores the advanced computational approach used to model hub protein interfaces from hot spots based on past studies, which provides the basis for work on hub protein function in PPINs. In current literature, the techniques depend largely on hub protein topology structure and gene expression, although we are mainly concerned with hub protein hot spots.

In this paper we present the feature selection approach that combines Pearson correlation coefficient with SVM-based recursive function elimination (SVM-RFE) to select a suitable subset of features. Instead, the XGBoost and LightGBM are generated based on structure functions to distinguish between hot spots and non-hot spots, Date-hub protein interfaces and Party-hub protein interfaces

II. PROCEDURE FOR PAPER SUBMISSION

A. Correlation-Based Feature Selection

Attribute selection is a process of selecting a subset of appropriate feature for model development to escape the curse of dimensionality which increases by over fit generalization [65]. Although certain design specifics may be overlooked, the features chosen are more descriptive. Under few examples, if the classifier is worked with a huge amount of functionality, the calculation cost is excessively huge and the grouping effectiveness is low. The examples in high-dimensional space might be changed by mapping or change to low-dimensional space. The out of date and incongruent capacities will be excluded through function collection to decrease the dimensionality.

The original datasets contain 59 characteristics: PSAIA acquired 36 structural attributes, 13 functional changes in monomers and complexes, and 10 physicochemical properties of 20 amino acids. A productive feature set is consisted of features thatare close ly linked to the class, but not associated to each other. The sub-set of capability must be powerful and flexible to accomplish the smallest sub-set of highlights which should not substantially corrupt the quality of the structure which impact class dissemination. We used the methodology of selecting highlights depending on the relationship coefficient to select highlights at that point, and we killed superfluous highlights by evaluating the vector for the similarity of highlights.

The Pearson relationship coefficient (PCC) is the most ideal approach to assist you with understanding the association among attributes and factors of response. The outcomes interim is [-1, 1], and - 1 speaks to a total negative relationship (A variable will ascend as the other one abatements), and + 1 speaks to a total positive connection, and 0 doesn't speak to a straight relationship. The paper uses the Pearson correlation coefficient as assess ment standard to assess the highlights, and aims to discover the exceptionally related highlights and evacuate the excess highlights.

B. Boosting Algorithm

It is a learning algorithm based on learning theory, which can develop a strong combination classifier with greater accuracy from a weak classifier system with low classification precision. The impact is more apparent particularly for less detailed classifiers, such as the decision tree.

The algorithm-boosting training method is ladder-like. Throughout every addition, it creates a new classifier. To determine the value of each sample, the classifier is used to label all the samples. Each time, weights are enhanced with misclassification from previous samples. Finally, a reliable and improved model of classification efficiency is obtained. The concept is expressed in Figure 3. An advanced algorithm based on conventional boosting algorithm, the gradient boosting (G-boosting) algorithm. G-boosting is distinct from the standard boosting method, which demonstrates improved learning performance. This creates the model for an iterative process like boosting but by reducing the loss function, it extends the model.

C. XGBoost

Extreme Gradient Boosting Algorithm referred to as XGBoost is a Machine Learning algorithm based on a decision-tree collection which uses a gradient boosting system. When solving issues affecting unstructured data, artificial neural networks appear to outperform all other algorithms or frameworks (images, text, etc.). XGBoost is a distributed gradient boosting library which is highly effective, scalable and portable. Machine learning algorithms are implemented utilizing the Gradient Boosting Architecture. XGBoost gives the tree (also known as GBDT, GBM) a parallel boost which addresses several data science issues quickly and effectively. The same technology operates on big distributed systems (Hadoop, SGE, MPI), and can solve problems in excess of trillions of cases.

D. LightGBM

It is a gradient boosting framework which uses node learning algorithms. It is planned to be delivered with the following advantages and is effective:

- Requires lesser memory
- Accuracy is enhanced
- Parallel and GPU learning is supported
- Can handle large volumes of data

III. RESULT AND DISCUSSION

As there are less reported findings regarding hub protein

hot spots we first used specific datasets to test our methods. We conducted studies using set of data from ASEdb, SKEMPI, and BID repositories. The SKEMPI collection comprises of 3047 binding-free energy shifts, which are obtained from the current data from 85 protein complexes. There are 485 traces, 1136 hot spots and 349 non-hot spots in the dataset. We have developed a separate training model for XGBoost and LightGBM based on SKEMPI dataset. The computational model using XGBoost provided a prediction accuracy of 99.8%. The model using LightGBM algorithm was ran for multiple times with different learning rates and each time the RMSE values were noted and a graph is plotted as shown below.

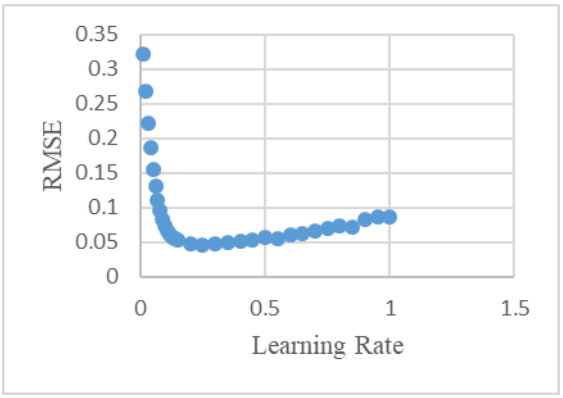

Fig. 1. Learning Rate v/s RMSE.

It can be seen from the figure 3.1 that when the learning rate is 0.25, the RMSE is minimum and we get the prediction accuracy as 99.95%.

IV. CONCLUSION

It is important to pick relevant features for the training within the small number of samples. The feature selection principle is for offering the smallest subset of attributes, without increasing the precision of the classification. To boost classification accuracy, the Pearson correlation coefficient is used to distinguish the more associated characteristics and delete redundant characteristics. Finally, for locating the hot spots of the different datasets and hub protein interfaces, extreme gradient boosting and LightGBM algorithms are used. LightGBM is found to outperform XGBoost.

REFERENCES

1. L. Giot, J. S. Bader, and et al.,"A protein interaction map of drosophila melanogaster," Science, vol. 302, no. 5651, pp. 1727-1736, 2003
2. P. Uetz, L. Giot, and et al., "A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae," Nature, vol. 403, no. 6770, pp. 623-627,2000.
3. A. J. Enright, I. Iliopoulos, and N. C. Kyrpides, Protein interaction maps for complete genomes based on gene fusion events, "Nature, vol. 402, no. 6757, pp. 86-90, 1999.
4. O. Keskin, A. Gursoy, and et al., "Principles of protein-protein interactions: what are the preferred ways for proteins to interact?" Chemical Rev., vol. 108, no. 4, pp. 1225-1244, 2008.
5. A. Pandey and M. Mann, "Proteomics to study genes and genomes," Nature, vol. 405, no. 6788, pp.837-846, 2000.
6. X. L. Lin, X.L. Zhang, and F.L. Zhou, "Protein structure prediction with local adjust tabu search algorithm," BMC Bioinf., vol. 15, no. S15, S1, 2014.
7. D. S. Huang, X. M. Zhao, and et al., "Classifying protein sequences using hydropathy blocks," Pattern Recognition, vol. 39, no. 12, pp. 2293-2300, 2006
8. D. S. Huang and X. Huang, "Improved performance in protein secondary structure prediction by combining multiple predictions," Protein and Peptide Letters, vol.13, no.10, pp. 985-991, 2006.
9. D. S. Huang and H. J. Yu, "Normalized feature vectors: a novel alignmentfree sequence comparison method based on the numbers of adjacent amino acids," IEEE/ACM Trans. Comput. Biology Bioinf., vol.10, no.2, pp.457-467, 2013.

AUTHORS PROFILE

Minul Vijayakumar received Bachelor of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering Kothamangalam in 2015 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. His research interest is in Bioinformatics and Machine Learning.

Joby George is currently an Associate Professor and Head of the Department of Computer Science and Engineering of Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. He received his B-Tech Degree in Computer Science and Engineering from Mahatma Gandhi University in 1994 and M-Tech in Computer Science and Engineering from IIT Bombay in 2005. His research interests include Bioinformatics and Machine Learning.