ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

An Efficient Technique for Protein Sequence Classification Using Data Mining

[1] Md. Arif Rahman, [2] Md. Alam Hossain, [3] Nazmun Nahar, [4] Must. Reshma Sultana
[1,2,3,4] Dept. of Computer Science and Engineering,
[1,2,3,4] Jessore University of Science and Technology,
[1,2,3,4] Jessore 7408, Bangladesh.

Abstract: -

Various classification techniques have been developed for the classification of protein sequences using feature selection method. Feature selection is important in accurate classification. This paper discussed some popular methods that involve feature selection for precise classification of protein sequences such as neural network based classifier, fuzzy method based classifier and rough set based classifier with their respective accuracy and drawback. Feature selection method is used for accurate classification of protein sequence. A new classification technique is proposed here following these discussed methods. The newness of the model proposed here is the aggregation of using intelligent method and introduction of a new technique for selecting specific features to classify protein sequence accurately and faster. Fuzzy classifier, neural method and rough set classifier are aggregated in a single model. The method's primary aim is to identify and classify the protein sequences based on extracting definite features from each sequence very fast with maintaining the accuracy level. Use of a new technique for extracting specific features reduces the computational overhead effectively. Comparing with the pervious methods a Great reduction of execution time without affecting accuracy level is achieved.

Keywords: - Data Mining; Protein Classification; Neural Network ; Fuzzy ARTMAP; Rough Set Classifier; 4-gram encoding method; 6-letter exchange group method.

1. INTRODUCTION

In twenty first century the world has entered in to a new epoch of information technology. New technology such as computer, mass storage, media etc. need to store a large quantity of data. Data mining has fascinated a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent prerequisite for turning such data into advantageous information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data mining can be viewed as a result of the natural fruition of information technology. For illustration, the early development of data collection and database creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. With numerous database systems offering query and transaction processing as common practice, advanced data analysis has unsurprisingly become the next aim.

To identify a particular character classification is important. With accordance to established criteria classification is the methodical arrangement in to groups. In present days exposing new biological information is a great challenge.

To conquer this challenge, we developed a new approach for protein classification. To identify protein classification is fundamentally important to store the variety of the known protein world. To classify protein sequence different classification technique has been proposed till today. Generally the Protein sequence is the arrangement of twenty different amino acids in some specific order.

Popular techniques of protein sequence classification depend on feature extraction from the sequence. Features are mainly the structural properties and also the functional properties of amino acids.

A comparison between the extracted features and the predefined values of those features are done due to classify the sequence. There are many eminent classification techniques such as neural networks, Genetic algorithm, FUZZY ARTMAP, ROUGH SET Classifier etc. None of the classification technique achieved 100% accuracy till date. A comprehensive study of ongoing research and a comparative analysis is presented in this paper.

The rest of the paper is structured as follows. Section II delivers an ephemeral review of the related techniques for classification of protein. Section III deliberates about our projected technique. Section IV provides the result and discussion achieved from our projected technique. Finally conclusions are provided.

2. CURRENT STATE OF ART

A. Neural Netwok

In [1][2][7][9] researchers used Neural Network model for protein sequence classification. In [1] researchers used n-gram encoding method to extract feature from the classifying protein sequence. The extracted feature than used to create pattern matrix. The accuracy level gained by this method is 90%. In [2] local and global similarities are extracted and matched with predefined values. [7] is the advanced technique of [1]. Here n-gram encoding method is used one by one. For selecting feature amino acid distribution, 2-gram amino distribution is used. SOM Network is used here. In [9] back propagation Neural Network is used. Extreme learning machine is used for classification purpose. Neural Network is generally good for non-linear data. Protein sequence is linear data so there is not much benefit of using this model for protein sequence.

B. Fuzzy ARTMAP

Fuzzy ARTMAP is a machine learning method does data by data analysis by calculating membership values. In [5][6][8] researchers used this method for classification. In [5] applying this method authors achieved 93% accuracy. Molecular weight (W), isoelectric point (IP), hydropathy distribution (D), Hydropathy Transformation (T) is calculated as features here. Authors of [6] calculate membership value as it is the most important in fuzzy model. For feature selection 6-letter exchange group method is used. Using the membership value pattern matrix is constructed. In [8] advancement of [6] is proposed. Here importance is given to reduce CPU time. Fuzzy ARTMAP model are time consuming because it does data by data analysis. Physical relationship cannot process with this model. Computational complexity is high. It also expansive model and has space complexity high.

C. Rough Set Classifier

In [10] [11] Rough set classifier is used for classification purpose. Rough set classifier has the abilities to overcome the disadvantages of the above classification mentioned above. In [10] rough set classifier a machine learning method is used which provide 97.7% accuracy. Rough set classifier generally consists of sequence arithmetic, Rough set theory concept lattice. Rough set classifier only gives knowledge based information. Data by data analysis is not conducted here. To classify a sequence in to a class or subclass this technique takes extra time and space.

D. New Classification Technique Using Neural Network, Rough Set, Fuzzy ARTMAP

In [23] author use a combined method including neural Network, Fuzzy ARTMAP and Rough set classifier for classification of protein sequence. This method proposed here use 3 segment for protein sequence classification. In Phase1 Fuzzy ARTMAP is used. For feature extraction molecular weight, isoelectric point, hydropathy distribution, isoelectric point, hydropathy composition, hydropathy transmission is calculated. If it fails to classify reduces dataset for 2nd phase. Phase2 implements Neural Network and work on reduced dataset collected from phase1. 2-gram encoding method and 6-letter exchange grouped method is used to extract feature from protein sequence. A pattern matrix is constructed with the feature value. Here back propagation Neural Network is used. Phase2 works when phase1 abortive to classify protein sequence. If both phase1 and phase 2 fail to classify sequence then phase3 works on the reduced dataset collected from the phase2. In phase3 neighborhood analysis is used. To use neighborhood analysis association rule is applied. It has supremacy to extract meticulous association between the protein sequence and classes, subclasses and families [23].This paper gives 98.5% accuracy. The noise elimination algorithm used in this paper is not strong. Accuracy level has opportunity to enhance and computational operating cost can be decreased.

E. Protein Sequences Classification Based on String Weighting Scheme

This paper presents a new technique for classification that is a combination of probabilistic modeling and supervised learning in a high dimensional feature space. This technique constructs feature vectors using Hidden Markov Model. The SVM classifier is used to detect the boundary value between two consecutive protein sequences. The proposed model takes in to account the conserved and non-conserved regions which is absent in the other previous methods for classification. The proposed model divides the protein sequence in to fixed length string subsequences. A HMM is used to assign a score to any sequence. If the two protein subsequences have the same k-length subsequences their HMM score will be close to each other. The scores used to search the entire database and to recognize and classify sequences. SVM use a set of training data to build a classifier by weighting each training item by how much it contributes to the overall classifier. The classifier built relies upon comparison between the weighted training data and the item which is to be classified. Here a data by data analysis occurred. The string weighting method for SVM significantly improves the classification protein domain based on remote homologies.

The main drawback of this method is it consider conserved and non-conserved regions within the protein sequences, the non-conserved regions score very low and if the sequence contains much of these regions it is problematic for the SVM classifier to achieve good accuracy.

3. PROJECTED TECHNIQUE

The function of this deliberate model is to categorize the unidentified protein sequence in to various families. This model also projected a high accuracy with a low computational time.

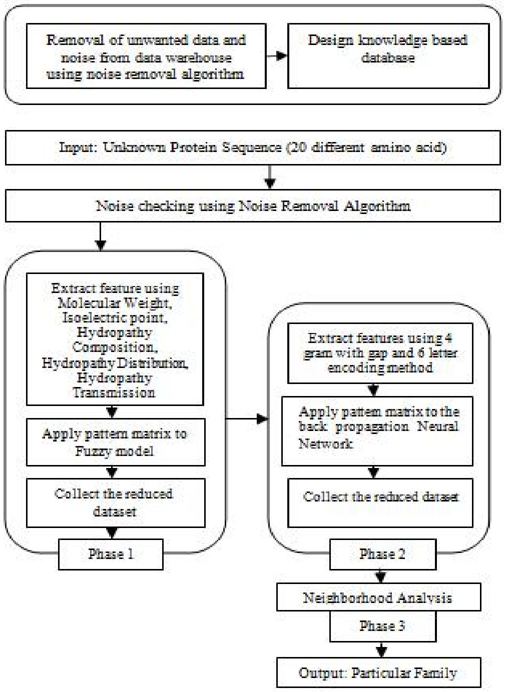
Removal of unwanted data and noise from data warehouse using noise removal algorithm
Design knowledge based database
Input: Unknown Protein Sequence (20 different amino acid)
Noise checking using Noise Removal Algorithm
Extract feature using Molecular Weight, Isoelectric point, Hydropathy Composition, Hydropathy Distribution, Hydropathy Transmission
Apply pattern matrix to Fuzzy model
Collect the reduced dataset
Phase 1
Extract features using 4 gram with gap and 6 letter encoding method
Apply pattern matrix to the back propagation Neural Network
Collect the reduced dataset
Phase 2
Neighborhood Analysis
Phase 3
Output: Particular Family

Figure . 1 Graphical representation of the proposed technique

To implement this purpose an unknown protein sequence is taken as an input. With the extraction of various features in singular phase and matching those with predefined values respected families are gained. For classify an unknown protein sequence feature selection plays a vital role here in this model.The model mainly works in three phase. First of all a cleaning process is made to the data warehouse with the help of noise removal algorithm. Then a knowledge based database is constructed from the values of data warehouse.

An unidentified protein sequence is taken as an input to this proposed model then a noise checking is performed to the input sequence using noise removal algorithm. If the input sequence is noise free then phase 1 is executed which helps to reduce the dataset for phase 2 input thus helps reducing complexity. Phase 2 helps to increase accuracy level without increasing CPU time for computation. Phase 3 implies the association rule in Neighborhood analysis method. All the three phases are capable of classifying the input protein sequence in to their respected families individually. If any one of the phase classifies the sequence in to its family the rest of the phase need not to execute. This helps diminution of complexity of this algorithm.

Phase 1**:** Fuzzy ARTMAP applied in phase1. Global feature is extracted for this model. Physic-chemical properties are considered here like molecular weight, isoelectric point, hydropathy distribution, hydropathy composition. Feature ranking algorithm is applied after extracting features. Fuzzy ARTMAP deals with data by data analysis and the features extracted here based on the molecular structure so a huge number of data can be reduced. Input protein sequence is classified in this segment otherwise data is reduced for segment2 which will execute next.

Phase 2: Neural Network model is used in phase 2. Knowledge based information is extracted here. For feature extraction a new technique is proposed. The input sequence consists of 20 different amino acids. In gram encoding model various pattern are selected for extracting features. Values of n those are too large, however will lack an ability to generalize beyond the training data because n-grams observed in the test sequence are unlikely to be observed in the training data. Likewise a value of n that is too short will be unable to effectively learn discriminatory features in the training data [24]. So choose 4 gram encoding method. Choosing 4gram required so many memory size and increase CPU time and cost. To deal with this problem gap technique is chosen which constructing pattern using 4gram encoding method.

For example consider a short sequence PVKTNKRPVKTNK.

4gram pattern are PVKT, VKTN, KTNK, TNKR, NKRP, KRPV, RPVK, PVKT, VKTN, KTNK. In the proposed technique the patterns will be PVKT, TNKR, RPVK, KTNK. Constructing patterns like this memory allocation will be decrease and also reduction of CPU time will occurred. A 6 letter exchange group method then applied for selecting features. After calculating pattern for every patterns value is calculated using as $\mathbf{x = c / (len(S) - 1}$ where x is the value, c is no of occurrence of every pair and Len(s) is the length of input sequence. These values are used to calculate mean value (m) and standard deviation (SD) using the formula:

$$m = (\sum^{N}_{i=1} x_i)/N, \quad d = \surd\ (\sum^{N}_{i-1}(X_i-m)\ ^{2}))/\surd\ (N-1)$$

These values are used as the input of Neural Network model.

Phase 3: To classify the input protein sequence in a particular family Neighborhood analysis will be used in the 3rd phase. Association rule generally use in the Neighborhood analysis. This rule can eliminate all other classes, subclasses and families of protein which the input sequence do not belongs it is possible because of having a power to extract the particular association between protein sequence and classes, sub classes and families.

In this method neighborhood analysis extracts the localized information. In the Neighborhood Association matrix least frequent occurring AA as center is computed for each AA for varying distances'd'. By considering binarization threshold 'T' the Binary Association Matrixes constructed. 0.05 And 5 are recommended as experimental result of 'T' and neighborhood distances'd'. Using Redacts based Decision Tree predominant attribute set is derived by attributing class information to the Binary Association Matrix to treating as a decision tree. For the identification of subclass a decision tree is constructed using predominant attribute values. The questioned protein will classify into a subclass using decision tree based Binary association Matrix for a family or for a class. For constructing concept lattice Binary Association Matrix is further used and attaching the given unfamiliar protein sequence to a set of protein.

4. RESULT AND DISCUSSION

To evaluate performance of the model proposed, a series of testing has been undertaken. Protein sequences and their consequent family name are obtained from NCBI and SCOP database. First, with the help of classification contrivance, protein family name and their corresponding values are stored in the database to create data warehouse. Then classifier contrivance is used to construct knowledge database and countrow database using data warehouse to reduce execution time. In the testing period five protein families are used that are FaeA-LIKE, Marine Metagenome Family WH1, MiaE-LIKE, PRP4-LIKE, and SOCS_BOX-LIKE. Testing is accomplished using 453 sequences as well as with various length including five families by this classifier tool. Within these 453 sequences, accurate results are found for all along a low time with the help of new feature selection technique. Three major techniques are used in this classification such as Fuzzy ARTMAP, Neural Network and Rough set classifier. Most of the sequences are classified at the first segment in Fuzzy model. The sequences which are failed to classify in 1st phase need 2nd phase that comprised of Neural Network for their classification. In Neural Network, a new technique is applied to reduce execution time and space. With the increasing value of n in n-gram encoding feature selection method, the computational cost also increases. So an n-gram encoding method with gap is proposed to minimize this cost.

After comparing between 4 gram encoding method and 4-gram gap method, it is found that the second method gives the best result and reduces a huge amount of time. So this overall process for classifing protein sequences is greatly acceptable. This new proposed feature selection technique in n-gram encoding method would be very helpful to other classification techniques and it would also be a better addition to this existing method.

The table paints the execution time of 4 gram encoding method and 4 gram encoding method with gap. Here from 5 different families, one protein sequence is executed using the classifier tool. From the table, it is seen that using 4 gram encoding method with gap, a great reduction of execution time is obtained.

Family Name	Execution time		Segment need
	Using 4-gram encoding method	Using 4-gram encoding methods with 3 letter gap	
PRP4-LIKE	1366	479	3rd
SOCS_BOX-LIKE	2151	492	2nd
Marine Metagenome Family WH1	5141	663	2nd
MiaE-LIKE	1045	269	2nd
FaeA-LIKE	1897	370	2nd

Table 1. PERFORMANCE COMPARISON BETWEEN TWO FEATURE SELECTIONS ENCODING METHOD.

A graphical representation of time comparison is shown in the chart. The chart depicts a great reduction of execution time.

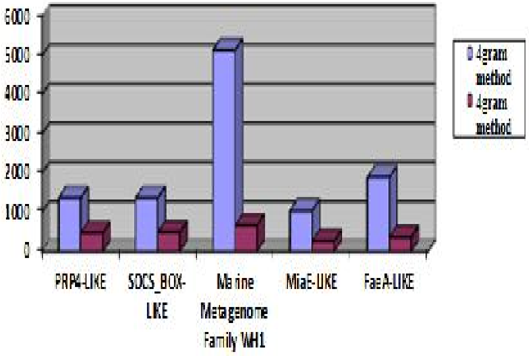

Figure. 2 Time Comparison

CONCLUSION

Data mining technology is capable of conducting large amount of data. Because of increasing data in organic field it is intellectual to use data mining to manage this bulky quantity of data. Different method is obtainable for categorization of protein sequence. This paper presents appraise of these techniques also shows that none of the obtainable techniques able to achieve 100% accuracy level. Here we projected a new feature extraction method for neural network model which is capable of extracting features within low CPU time and space. Here a classification technique is proposed consist of three methods the Neural Network, Rough set classifier, fuzzy ARTMAP model. With this proposed model reduction of CPU time, space and a high accuracy is gained. Here used methods such as Fuzzy ARTMAP, Neural Network are much customized and the noise removal algorithm is also strong for detecting noisy data. In future more analysis and studies will be done for the enhancement of this classification system.

REFERENCES

[1] V. V. Solovyev and K. S. Makarova. (1993) A novel method of protein sequence classification based on oligopeptide frequency analysis and its application to search for functional sites and to domain localization, Computer Applications in the Biosciences, 9(1): 17-24.

[2] C. Wu, M. Berry, S. Shivakumar , J. Mclarty (1995) Neural Networks for Full-Scale Protein Sequence Classification: Sequence Encoding with Singular Value Decomposition. Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, Machine Learning, 21, pp: 177-193.

[3] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. Kim (September, 1995) Prediction ofprotein folding class using global description of amino acid sequence,in Biophysics, vol. 92, (USA), pp. 8700–8704, National Academy of Science, USA.

[4] C. H. Wu and J. McLarty (2000) Neural Networks and Genome Informatics. Elsevier Science.

[5] J. T. L. Wang, Q.H. Ma, D. Shasha, C. H Wu (2000) Application of Neural Networks

to Biological Data Mining: A case study in Protein Sequence Classification. KDD, Boston, MA, USA, pp: 305-309.

[6] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Research, vol. 31, pp. 3692–3697.

[7] D. Wang, G.B. Huang (2005) Protein Sequence Classification Using Extreme Learning Machine. Proceedings of International Joint Conference on Neural Networks (IJCNN2005), Montreal, Canada.

[8] X. Zhao, Y. Cheung, and D. Huang (October 2005) A novel approach to extractingfeatures from motif content and protein composition for proteinsequence classification, Neural Networks, vol. 18, pp. 1019–1028.

[9] S. Mohamed, D. Rubin and T.Marwala (2006) Multi-class Protein Sequence Classification Using Fuzzy ARTMAP. IEEE Conference pp: 1676 – 1680.

[10] E. G. Mansoori, M. J. Zolghadri, S. D. Katebi, H. Mohabatkar, R. Boostani and M. H. Sadreddini (2008) Generating Fuzzy Rules For Protein Classification. Iranian Journal of Fuzzy Systems Vol. 5, No. 2**,** pp. 21-33.

[11] Z.Zainuddin, M. Kumar (2008) Radial Basic Function Neural Networks in Protein Sequence Classification. Malaysian Journal of Mathematical Science2(2), pp: 195-204.

[12]E. G. Mansoori, M. J. Zolghadri, and S. D. Katebi(March 2009) Protein Superfamily Classification Using Fuzzy Rule-Based Classifier. IEEE Transactions OnNanobioscience, Vol. 8, No. 1, pp 92-99.

[13] PV NageswaraRao, T Uma Devi, DsvgkKaladhar,Gr Sridhar, AllamAppaRao (2009) A Probabilistic Neural Network Approach For Protein Superfamily Classification. Journal of Theoretical and Applied Information Technology.

[14] R.Yellasiri, C.R.Rao (2009) Rough Set Protein Classifier. Journal of Theoretical and Applied Information Technology.

[15] S. A. Rahman, A. A. Bakar, Z.A. Mohamed Hussein (2009) Feature Selection and Classification of Protein Subfamilies Using Rough Sets. International Conference on Electrical Engineering and Informatics, Selangor, Malaysia.

[16] K.Boujenfa, N. Essoussi, M. Limam, (2011) Tree-kNN: A Tree-Based Algorithm for Protein Sequence Classification” Vol. 3, PP: 961 – 968, ISSN: 0975-3397, International Journal on Computer Science and Engineering (IJCSE).

[17] P. G. Ferreira and P. J. Azevedo, (2010) Protein Sequence Classification through Relevant Sequence Mining and Bayes Classifiers.

[18] R. Busa-Fekete, A.Kocsor ,S.Pongor, (2010) Tree-Based Algorithms for Protein Classifcation. International Journal on Computer Science and Engineering (IJCSE).

[19] G. Tzanis, C. Berberidis, and I.Vlahavas, “Biological Data Mining”.

[20] J. Han, M.Kamber, (2001) “Data Mining: Concepts and Techniques” Morgan Kaufmann Publishers, 1st Edition.

[21] M. H. Dunham, (2006) “Data Mining: Introductory and Advanced Topics”, Pearson Education, 1st Edition.

[22] S. Saha, R. Chaki (2012) “A Brief Review of Data Mining Application Involving Protein Sequence Classification”, Advances in Intelligent Systems and Computing, 1, Volume 1771 Advances in Computing and Information Technology, Springer Publisher, ACITY 2012, Chennai, India, pp. 469-477.

[23] S. Saha, R. Chaki (2012) “Application of Data Mining in Protein Sequence Classification”, International Journal of Database Management System (IJDMS), Volume 4, Number 5, October 2012, AIRCC, Springer Publication, pp. 103-118,

ISSN: 0975-5705 (Online), 0975-59851(Print).

[24] B. R. King (2008) ”Protein Sequence Classification with Bayesian Supervised and Semi-Supervised Learned Classifiers”.

[25] M. J. Iqbal, I. Faye, B. B. Samir, and A. Md Said (2014) “Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics” , The Scientific World Journal Volume 2014 (2014), Article ID 173869

[26] Dr.S. Vijayarani and Ms. S.Deepa(2014)” Protein sequence Classification In Data Mining- A Study” International journal of information technology ,Modeling and Computing(IJITMC),Vol.2,No.2,May 2014.