


Chapter 1

A Biological Data–Driven Mining Technique by Using Hybrid Classifiers With Rough Set

Linkon Chowdhury

 <https://orcid.org/0000-0002-6345-4670>
East Delta University, Bangladesh

Md Sarwar Kamal

University of Technology Sydney, Australia

Shamim H. Ripon

East West University, Bangladesh

Sazia Parvin

University of New South Wales, Australia

Omar Khadeer Hussain

University of New South Wales, Australia

Amira Ashour

Tanta University, Egypt

Bristy Roy Chowdhury

BGC Trust University, Bangladesh

ABSTRACT

Biological data classification and analysis are significant for living organs. A biological data classification is an approach that classifies the organs into a particular group based on their features and characteristics. The objective of this paper is to establish a hybrid approach with naïve Bayes, apriori algorithm, and KNN classifier that generates optimal classification rules for finding biological pattern matching. The authors create combined association rules by using naïve Bayes and apriori approach with a rough set for next sequence prediction. First, the large DNA sequence is reduced by using k-nearest approach. They apply association rules by using naïve Bayes and apriori approach for the next sequence pattern. The hybrid approach provides more accuracy than single classifier for biological sequence prediction. The optimized hybrid process needs less execution time for rule generation for massive biological data analysis. The results established that the hybrid approach generally outperforms the other association rule generation approach.

DOI: 10.4018/979-8-3693-3026-5.ch001

INTRODUCTION

Automated biological data analysis, processing, and synthesis are an active research domain. Massive amount of the biological data produced and stored continuously in biological studies. Information retrieving and analysis from these large sizes of generated data is a challenging issue as well as a critical factor for successful knowledge discovering. Analytical and mathematical approaches used to mine organized and unorganized biological data from various volumes of data. However, it is challenging to conquer hidden and unknown significant data from multiple databases irrespective to their sizes and positions. Hence, mining algorithms, techniques, processes, tools and frameworks are equally valuable for data acquisition and information mining. Recently, the accumulations of biological information and datasets have been rising dramatically due to the development of such tools. The rapid increase of bioinformatics and microbiological datasets appealed to the dependence of computers for the arrangement, computation and synthesis of the datasets. Among the different computational approaches, parallel processing is more effective in biological data operation (next-generation sequencing, genome analysis, etc.) and new data set generation (Kibegwa, F. M., Bett, R. C., Gachuri, C. K., Stomeo, F., & Mujibi, F. D. (2020)(2Li, C. X., Li, W., Zhou, J., Zhang, B., Feng, Y., Ping Xu, C., Lu, Y. Y., Holmes, E. C., & Shi, M. (2020)(3Sangphukieo, A., Laomettachit, T., & Ruengjitchatchawalya, M. (2020)(Kopf et al., 2015).

Diversity of human, animal and plant life/behavior is one of the primary reasons behind the availability of the biological datasets. Despite the variations, the biological functionalities interrelated with sets of living organs that are common in-universe. These organs solely depend on the different microbiological functionalities of proteins, DNA (Deoxyribonucleic acid), and RNA (ribonucleic acid). Microbiological/metagenomic data analysis processes simpler illustrations and sequencing for a metagenomic dataset for vast microbial data diversity (Villar et al., 2015), (Sunagawa et al., 2015). There are several types of metagenomic datasets that are dominant in biological data mining. For large dataset mining, there are large-scale genome sequences that require efficient representations and processing. Consequently, high-performance algorithms and techniques are demanding in biological data analysis and research (Lu et al., 2015)(Li et al., 2015)(De Cruz et al., 2015).

Classification is a significant approach for massive biological data analysis and mining. The classification system includes a sequential combination of multiple methods used to build an efficient solution to deal with particular and correlated data classification. Meta-learning classifiers are recently popular in genome researchers (Statnikov et al., 2005). Such learning classifiers refer to manipulate a set of base predictors for a given classification task and then correlate the output information by using an integration technique. Association rules or classification rules generated under different names such as classifier ensembles, consensus aggregation, decision combination, classifier fusion, hybrid methods and more (Kuncheva, 2002), (Dasarathy, 1994). To improve the performance of a single classifier, the priority assigned to the association or classification rules. Different types of classifiers usually make different predictions outcome on the same sample set of data. Data diversity illustrated that the sets of misclassified samples are classified from different uncorrelated sources by using multiple sets of classifiers. To classify various data diversity used for priority association table. The techniques to develop priority association table divided into two categories, namely, classifiers disturbance and sample disturbance. The first approach of the classifier utilizes the instability of the base classifiers. These classifiers are very sensitive to the initialization parameters. Such classifiers include machine learning, neural networks, random forests, and decision trees. The second approach of the classifiers trains the feature of data sets with different sample subsets.

18 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage:
www.igi-global.com/chapter/biological-data-driven-mining-technique/342519

Related Content

Human Oral Bioavailability Prediction of Four Kinds of Drugs

Aixia Yan, Zhi Wang, Jiaxuan Li and Meng Meng (2011). *Interdisciplinary Research and Applications in Bioinformatics, Computational Biology, and Environmental Sciences* (pp. 141-154).

www.irma-international.org/chapter/human-oral-bioavailability-prediction-four/48372

Pharmacoinformatics: Advanced Information Systems for Improved Pharmaceutical Care

Tagelsir Mohamed Gasmelseid (2012). *Pharmacoinformatics and Drug Discovery Technologies: Theories and Applications* (pp. 1-11).

www.irma-international.org/chapter/pharmacoinformatics-advanced-information-systems-improved/64062

Mitochondrial Pyruvate Carrier 1 and 2 Heterodimer, In Silico, Models of Plant and Human Complexes: A Comparison of Structure and Transporter Binding Properties

Jason L. Dugan, Allen K. Bourdon and Clyde F. Phelix (2017). *International Journal of Knowledge Discovery in Bioinformatics* (pp. 11-42).

www.irma-international.org/article/mitochondrial-pyruvate-carrier-1-and-2-heterodimer-in-silico-models-of-plant-and-human-complexes/190791

Ensemble Gene Selection

(2011). *Feature Selection and Ensemble Methods for Bioinformatics: Algorithmic Classification and Implementations* (pp. 329-333).

www.irma-international.org/chapter/ensemble-gene-selection/53911

Discovering Network Motifs in Protein Interaction Networks

Raymond Wan and Hiroshi Mamitsuka (2009). *Biological Data Mining in Protein Interaction Networks* (pp. 117-143).

www.irma-international.org/chapter/discovering-network-motifs-protein-interaction/5562