THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Fuzzy Inference System For Predicting Protein Functional Activity Using Data Mining

Sahil Sharma
School of Computer Science Engineering
Lovely Professional UniversityPhagwara,Punjab, India
Sahil.24886@lpu.co.in

ShikhaJary
Chandigarh University,Punjab, India
shikhajaryal57@gmail.com

ABSTRACT

The progress of bioinformatics generates a large volume of data that needs to be analyzed in order to identify various structures. Proteins, involved in several important tasks in aliving organism. arc a mixture of amino acids which have different structures and patterns. The number of primary structures solved and stored in databases is growing faster than our capability to solve these tertiary structures using different experimental methods and so is the error occurring in them is more. To have efficient and effective protein prediction we need to have accurate results that can be obtained by minimizing the errors in the proteins. The analysis of bioinformatics data has seen major shifts from traditional data mining approach to hybrid approaches. Fuzzy inference system is proposed with and relief algorithm for decreasing error rate in theprediction of proteomic data. The fuzzyrule helps to recognize the uncertainties and vagueness in patterns of the protein structure. The methodology considers the perception and cognitive uncertainty of subjective decisions allowing the usage of theimprecise description of protein data.

Keywords: Data Mining, Fuzzy Inference System

1. INTRODUCTION

It implies collecting or "mining" useful information from a ton of data. Data Mining is an examination of finding new interesting cases and association in the hugequantity of data. It is determined as "a way toward discovery important new connections, for examples, also patterns by delving into a lot of information put away in warehouses[1]. Data mining is additionally in some cases called Knowledge Discovery in Databases (KDD)[2]. Information mining is not limited to particular industry. It need carry out innovations and the excitement to inspect the possibility of covered information that dwells in the information. Data Mining

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

approaches have all the earmarks of being preferably suited for Bio Informatics because it is data rich, yet does not have an extensive speculation of life's relationship at the sub-nuclear level. The broad databases of organic data make both difficulties and open doors for improvement of novel KDD strategies. Mining natural information removes helpful learning from enormous data sets accumulated in science, and in other related life sciences regions, for example, pharmaceutical and neuroscience.

Data mining is a method of examininga huge amount of database to produce new information in it. It is the process to analyze the data and generate some important information in terms of rules or patterns. Data mining is an influential technology with the potential to help persons focus on most essential information in data warehouses. Different tools with techniques are used to extract patterns and information that are hidden in the large databases. These tools can predict future trends and behaviors and have the capability to answer desired questions that were traditionally too time-consuming to resolve. Data mining is an integral part of KDD. Data mining plays important tasks in the field of research and practical applications. The main challenges to the data mining are as follows

1. Huge datasets and high dimensionality
2. Understandability of the patterns
3. Redundant data
4. Incomplete data and data Integration

Above issues of data mining can be solved by various other techniques of data mining.

2. BIOINFORMATICS

With more natural data produced, the most squeezing assignment of Bioinformatics has gotten to be to investigate and decipher different sorts of information, including nucleotide and amino corrosive successions, protein structures, quality expression profiling and relevance information mining methods of highlight era, include determination, and highlight combination with learning calculations to handle the issues of illness phenotype grouping and patient survival forecast from quality expression profiles, and the issues of utilitarian site expectation from DNA arrangements[5].

Throughout late decades fast changes in genomic and other nuclear research developments and progressions in information propels have joined to convey an epic measure of information related to sub-nuclear science.

Fuzzy rules are the new approach that is used to mine quantitative data frequently present in

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

the databases. Fuzzy rules use the fuzzy logic to change numerical type attributes to fuzzy type attributes[13]. The fuzzy data mining technique that is Fuzzy Repeated Pattern growth (FRP-growth) algorithm can be used to mine the large datasets. It treats each data item as variable and its portioned is based on the linguistics values.

3. LITERATURE REVIEW

Stefano Cresci*et al(2016).* In this creator proposed a strikingly novel, basic, and powerful way to deal with model online client conduct, it separate and dissect computerized DNA successions from client online activities and utilize Twitter as a benchmark to test the proposal[15]. It additionally acquires a sharp and minimized DNA-roused portrayal of client activities. At that point, it applies standard DNA investigation strategies to separate amongst veritable and spambot accounts on Twitter. A test battle underpins this proposition, demonstrating its adequacy and reasonability. To the best of information, they guaranteed to be first ones to recognize and adjust DNA-propelled systems to online client behavioral displaying. While Twitter spambot discovery is a particular utilize case on a particular web-based social networking, they proposed approach is stage and innovation rationalist, thus making ready for various behavioral portrayal errands.

M. Masseroli*et al(2015).* In this examination paper noteworthy advances in biotechnology and structure science is done which is making an astounding measure of bimolecular data and semantic remarks, moreover which increases in number and quality, yet it is scattered and just for the most part related. Compromise and mining of these scattered and making information and data have the high limit of finding canvassed biomedical learning obliging in appreciation complex natural miracles, run of the mill or hypochondriac, and finally of enhancing investigation, foresight, and treatment, however such fuse stances gigantic challenges[16]. Also, it has endeavored to handle them by working up a novel and summed up way to deal with describe and successfully keep up overhauled and intensify a compromise of many progressing and heterogeneous data sources. Maker's approach showed handiness to expel biomedical data about complex natural methodology and diseases. Understanding complex normal marvels incorporates taking note of complex biomedical request on different biomolecular information at the same time, which are passed on through different genomic and proteomic semantic comments Scattered in many appropriated and heterogeneous information sources, such heterogeneity and scattering hamper the master's capacity of asking general request and performing general assessments. To beat this issue, producer has made programming planning to make and keep up a Genomic and Proteomic

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Knowledge Base, which combines two or three the most significant wellsprings of such scattered data

Alfredo Cuzzocrea*et al(2015).*The author found that Provenance of Big Data is an interesting issue in the database and information mining research groups[17]. Fundamentally, provenance is the way toward identifying the genealogy and the inference of information and information items, and it assumes a noteworthy part in database administration frameworks and also in work process administration frameworks and conveyed frameworks. Regardless of this, the provenance of huge information research is still in its embryonic stage, and a considerable measure of endeavors should at present be done in this area. Motivated by these contemplations, this paper gives an outline of pertinent issues and difficulties with regards to huge information provenance examine, by likewise highlighting conceivable future endeavors inside these examination headings.

Eric P. Xing*et al(2015).*If there is any precise approach to productively apply a large range of cutting edge ML projects to modern scale issues then that is, utilizing Huge Models ranging up to hundreds of billions of different parameters used in on Big Data which reaches up to terabytes or pet bytes in size[18]. Cutting edge parallelization procedures utilize fine-grained operations and booking past the great mass synchronous preparing worldview promoted by MapReduce, or even specific chart construct execution that depends with respect to a diagram which indicates the use of ML programs. The assortment within methodologies tries to describe frameworks and calculations outline in various headings, and it stays hard to locate an all inclusive stage pertinent to an extensive variety of ML projects at scale. A universally useful structure, Petuum ,that methodically addresses information and model-parallel difficulties in expansive scale ML, by watching that numerous ML projects are in a general sense enhancement driven and concede blunder tolerant, iterative-joined algorithmic arrangements. Creator exhibited the viability of these framework plans versus surely understood the usage of present day ML calculations, demonstrating that Petuum permits ML projects to keep running in substantially minimal time and with the impressively bigger representation sizes, or we can say even on unobtrusively measured process groups.

Suvarna*et al(*2015) In this examination paper critical advances in biotechnology and structure science is done which is making an amazing measure of biomolecular information and semantic comments, besides which increments in number and quality, yet it is scattered and only generally related[19]. The trade off and mining of these scattered and making data and information have the high furthest reaches of discovering peddled biomedical learning

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

obliging in gratefulness complex regular supernatural occurrences, common or despondent person, lastly of upgrading examination, premonition, and treatment, however such circuit positions monstrous challenges. Likewise, it has tried to handle them by working up a novel and summed up approach to managing to depict and effectively keep up upgraded and strengthen a bargain of many advancing and heterogeneous information sources. Producer's approach demonstrated handiness to oust biomedical information about complex normal system and illnesses. Understanding complex typical wonders consolidate observing complex biomedical demand on various biomolecular data in the meantime, which are gone on through various genomic and proteomic semantic remarks Scattered in many appropriated and heterogeneous data sources, such heterogeneity and dispersing hamper the ace's ability to ask general demand and performing general appraisals. To beat this issue, maker has made programming wanting to make and keep up a Genomic and Proteomic Knowledge Base, which consolidates a few the most noteworthy wellsprings of such scattered information

SanjimaManocha*et al(2014).* In this, the creator has investigated the essential components of information mining strategies in distributed computing and securing the information utilizing edge identification strategy. What's more, tries to incorporate information mining strategies into distributed computing and picture handling making it a mutt approach[20]. The use of data mining methodology through disseminated figuring inclinations the customers to separate important concealed prescient data from for all intents and purposes incorporated information stockroom that decreases the expenses of capacity and foundation. Unify administration of programming and information stockpiling, with an affirmation of effective, solid also, secure organizations for their customers through Edge area based approach for picture steganography. As the learning to be removed from the information is put away over the distributed storage, likewise, the trouble to get every one of the information to a brought together capacity is discussed. Subsequently, the significance for an information mining calculation that works over non incorporated information stockpiling over dispersed environment is examined.

SanjimaManocha*et al(2014).* In this, creator has investigated the essential components of information mining strategies in distributed computing and securing the information utilizing edge identification strategy. What's more, tries to incorporate information mining strategies into distributed computing and picture handling making it a half-breed approach[20]. The usage of information mining procedures through distributed computing urges the clients to separate important concealed prescient data from for all intents and purposes incorporated

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

information stockroom that decreases the expenses of capacity and foundation. Unify administration of programming and information stockpiling, with an affirmation of effective, solid and secure administrations for their clients through Edge location-based approach for picture steganography. As the learning to be removed from the information is put away over the distributed storage, likewise, the trouble to get every one of the information to a brought together capacity is discussed. Subsequently, the significance for an information mining calculation that works over non-incorporated information stockpiling overdispersed environment is examined.

AvitaKatal*et al(2013)* portrayed Big information as a significant measure of information which requires new improvements and models with the target that it finds the opportunity to be unmistakably conceivable to think respect from it by getting and examination handle. Because of such interminable size of information, it winds up being enormously hard to perform productive examination utilizing the present routine techniques[21]. Immense information because of its unmistakable properties like volume, speed, gathering, fluctuation, respect and multifaceted nature set forth various inconveniences. Since Big information is a late pending improvement in the market which can pass on immense ideal conditions to the business affiliations, it persuades the chance to be especially important that various inconveniences and issues related to passing on and changing as per this progression are gotten into light this paper. This paper likewise presented the Big information headway near to its significance in the present world and existing attempts which are persuading and essential in changing the likelihood of science into goliath science and society as well. The unmistakable difficulties and issues in adjusting and persevering Big information improvement, its mechanical congregations like Hadoop are in addition broke down in detail near to the issues Hadoop is going up against.

4. PROBLEM DEFINITION

Proteins are the inadequately comprehended and anticipated in view of proteomic and genomic information sets accessible. The expectation for the most part incorporates information serious computational mining. Robotized comment foreseeing protein work exercises are extremely mind boggling. With the expansion in the quantity of sequenced genomes quickly, all the protein items must be commented on numerically.

Conditions upon computational forecasts then it is vital that the estimation of these techniques be high. Different discoveries have appeared now and again that:

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

(i) Nowadays best protein work forecast calculations unmistakably beat all the more oftentimes utilized more established era techniques, with significant advantages on a wide range of target information and

(ii) Even the highest strategies have indicated alright learning to guide tests.

The right expectation of protein capacity is primary component to comprehend and see life at the point by point sub-atomic level and has high biomedical and pharmaceutical ramifications. Be that as it may, with this much-inbuilt intricacy and expense, the speculative portrayal of capacity can't extend to hold the tremendous measure of arrangement information effectively accessible. The computational clarification of protein capacity has subsequently risen as a difficulty at the cutting edge of computational and atomic science.

Various arrangements have been foreseen in the most recent four decades, however, the errand of computational utilitarian derivation as often as possibly depends on customary methodologies, for instance, recognizing spaces or discovering basic kept Alignment Search which is a principle in the midst of proteins with tentatively firm capacity.

The work presented in this paper is based on the Fuzzy Association rule to mine proteins data. Here we extract rules to associates patterns to specific secondary structures of proteins. Further, we have seen the error rate and have tried to minimize the error and therefore the main emphasis is to propose a fuzzy data mining technique to find the fuzzy association rules using fuzzy partition method and FP-growth on the proteins datasets.

Lately, the openness of genomic-level succession data for a large number of animal groups, tied with huge high-throughput investigational information, has made new open doors for the capacity forecast. An incredible number of strategies have been foreseen to use this information, including capacity expectation from amino corrosive grouping, protein-protein communication systems, and protein structure information.

An unbiased appraisal of these distinctive strategies can give understanding into their ability to separate proteins practically and can control natural investigations. In this way, however, a total assessment consolidating a substantial and divergent arrangement of target groupings has not been led due to down to earth troubles and issues in giving an effectively explained target set.

Data analysis along with data mining of all the data is a very staggering task due o the

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

inconsistency and vagueness in the data generated to uncover the uncertainties in the data fuzzy computations can be considered by effective evaluation of the proteomic data

5 ALGORITHM USED

On the available dataset the feature selection is applied to do preprocessing to identify the subsets of data that mainly compute the results from a large amount of data and provides us with the features or attributes that haveminimal and maximal contribution tocomputing results.

For this purpose, we have used the Relief algorithm which helps to improve the accuracy of classification and also reflects the attributes that need to be taken care of to improve the results.

5.1 Relief algorithm

Relief algorithm is the easy, high-speed, and efficient approach for attribute weighting. The end result of Relief algorithm is an assignment of weights which lie between the range of −1 to 1 for every single attribute, and with more number of positive weights shows or indicate more attributes that can be predicted[29]. Relief Algorithm has various variants which further depend on the category or nature of data also on attributes and characteristics of data. This Algorithm workings are based on the principles as follows:

In this, the weight of a given attribute is updated in different iterations in the procedure mentioned below. A random sample of data is selected, after that the sample is identified which lie in the same zone or belongs to the same given class and neighboring nearest sample is called (nearest hit) and on the other hand those samples which do not lie nearby and belong to some other different class are identified and are said to be belonging to opposite class which is termed as (nearest miss). Relief algorithm basically works on this principle of Nearest Hit and Nearest Miss which are building blocks of this algorithm. The next thing we calculate is the distance between Nearest Hit and Chosen Sample and same with Nearest Miss and chosen Sample which is calculated on the basis of Manhattan distance between two points. Then there is a change in feature weights which are taken into consideration for feature selection in the classification of the given target class. These features found are then given extra weights for the process of classification. So we can say that weights obtained of feature play very important role for locating the accurate or correct class.

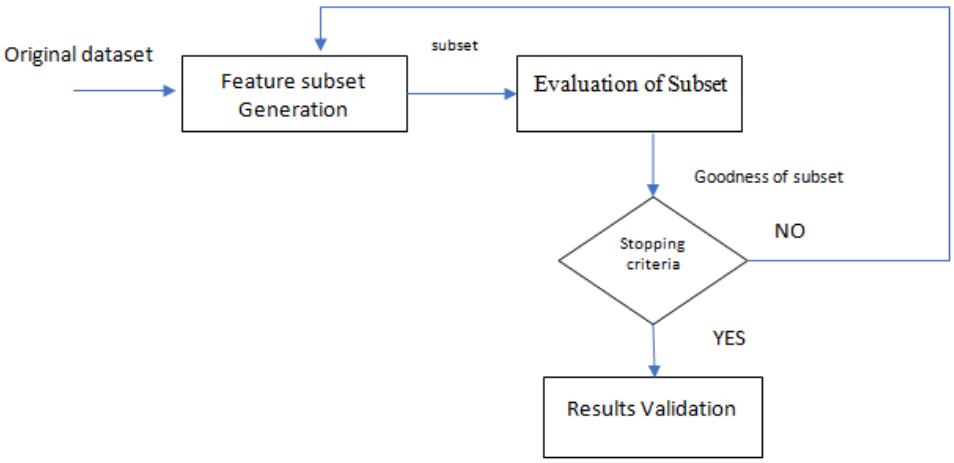
Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

The Relief algorithm used is shown as under which follows the steps as shown[30]:

Input: For every instance training data there are vector values of Attribute and Class which is called Class label.
Output: Estimated vector *W* with corresponding weights of different components.

1. Set all the weights W[A]:=0.0;
2. for every i:=1 to m do begin
3. Randomly select an instance r_i ;
4.Then find k-nearest hits h_j;
5. for every class C class(r_i) do
6. from class C find k-nearest misses m_j (c);
7. for A:=1 to a
8. W[A]=W[A]- $\sum_{j=1}^{k} \frac{diff\ (A,r_i,h_j)}{m.k}$ + (1)
9. $\sum_{C \neq class\ r_i} \frac{\left[\frac{p(c)}{1-p(class\ (r_i} \sum_{j=1}^{k} diff\ (a,r_i,h_j)\right]}{m.k}$ (2)
10. End



CONCLUSION

Various authors have tried to predict the protein structure and protein functional activates, so

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

we have the different techniques and methodologies available. But as with the passage of time the data required to be processed has increased manifold so we need to focus on the accuracy of the results predicted. For that the error rate that occurred in prediction need to be minimized. Because not only in research areas proteins are quite helpful in other aspects also like to invent new drugs, diagnose the patients with diseases. So in this key areas not only prediction will solve our purpose, we need to reduce or minimize the average error rate as well so that we can be more sure about the predicted end results.

As discussed earlier that the error rate which was given by other authors on the yeast dataset is 0.36806 which is minimum as per their study. On the other hand we have tried and successfully reduced the average error rate on the same dataset and the new improved results show that we have 0.35945 average error rate.

FUTURE SCOPE

Fuzzy method can be used to efficiently analyze the Bioinformatics. Some of the fuzzy applications are intelligent control and pattern classification which can be incorporated for structuring the proteomic data. This method help different user to prepare their knowledge and to simplify the procedure to recover and explore the proteomic information in a most simple way. Association mining is used to express the relationships among the objects and can be used to mine the proteomics data from bioinformatics.

So the improvement in the average error rate using Relief algorithm over the previous existing technique is concluded in the given table below.

Table 1: Comparison of Error rate in different Trapezoidal Curves

Description	**Input Membership Function**	**Output Membership Function**	**Results**
Existing error rate	Trapezoidal	Triangular	0.36806
New Improved error rate	Trapezoidal	Triangular	0.35945

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Various extraction and analysis techniques are also discussed for mining and analyzing bioinformatics data. Bioinformatics is a vast topic in which various research work are done by researchers. For mining proteomics data using fuzzy association rule can be used

Usages of data mining to Bioinformatics fuse quality finding, protein work space acknowledgment, work subject distinguishing proof, protein work inference, disease investigation, contamination reckoning, affliction treatment headway, protein and quality affiliation sort out proliferation, data purification, and protein cells region estimate.

REFERENCES

[1] J. L. C. Ramos, R. E. D. Silva, R. L. Rodrigues, J. C. S. Silva, and A. S. Gomes, "A Comparative Study between Clustering Methods in Educational Data Mining," vol. 14, no. 8, 2016.

[2] B. E. V. Comendador, L. W. Rabago, and B. T. Tanguilig, "An educational model based on Knowledge Discovery in Databases (KDD) to predict learner's behavior using classification techniques," *2016 IEEE Int. Conf. Signal Process. Commun. Comput.*, pp. 1–6, 2016.

[3] J. P. A. Ioannidis, "Why most published research findings are false: Author's reply to Goodman and Greenland [7]," *PLoS Medicine*, vol. 4, no. 6. pp. 1132–1133, 2007.

[4] D. M. Creasyand B. T. Tanguilig,"Bioinformatics, Data ming tasks vol. 17, no. 24, pp. 3481–3571, 2014" .

[5] "Guest Editorial : Data Mining in Bioinformatics , Biomedicine , and Healthcare Informatics," vol. 18, no. 2, p. 2306988, 2014.

[6] L.Atzori, A. lera, G. Morabito, The Informatics in Biology: Survey. Computing through different ways , 2787– 2805. 2018

[7] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, no. 18, pp. 3551–3567, 1999.

[8] Theoleyre, F., & Pang, A. C. (Eds.). Internet of Things and M2M Communications. River PubliC.20(6),71-78. (2016)

[9] "protein-structure-prediction-using-coarse-grain-force-fields-6530787 @ www.slideshare.net." .

[10] J.-K. Yu, Y.-D. Chen, and S. Zheng, "An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics.," *World J. Gastroenterol.*, vol. 10, no. 21, pp. 3127–31, 2004.

[11] S. Mission, S. Mission, H. Topics, and H. Topics, "The Eighth Annual Bio-Ontologies Meeting," *PLoS Comput. Biol.*, vol. 1, pp. 2005–2006, 2005.

Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[12] A. Rai, "Concepts of Bioinformatics," *Online Content Creat. Manag. an eLearning Environ.*, pp. 333–352, 2001.

[13] N. H. Reyes, "Fuzzy Inference Systems," 2012.

[14] "node16 @ www.cs.bris.ac.uk." .

[15] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "DNA-inspired online behavioral modeling and its application to spambot detection," 2016.

[16] A. Cuzzocrea, "Provenance Research Issues and Challenges in the Big Data Era," pp. 684–686, 2015.

[17] alfredoMann. MNachman, 1., & Pe'er, D. (2000).Issues , Challenges , Tools and Good Practices in " vol. 5963, no. c,7, pp. 7–11, 2013

[18] E. P. Xing *et al.*, "Petuum : A New Platform for Distributed Machine Learning on Big Data," vol. 1, no. 2, pp. 49–67, 2015.

[19] S. V. Koneru and D. B. S, "Divide and Conquer Approach to Contact Map Overlap Problem using 2D-Pattern Mining of Protein Contact Networks," vol. 5963, no. c, pp. 1–9, 2015.

[20] S. Manocha, "A Novel Hybrid Approach for Secure Cloud Mining using Lossless Image Format," vol. 98, no. 7, pp. 7–11, 2014.

[21] A. Katal, "Big Data : Issues , Challenges , Tools and Good Practices," pp. 404–409, 2013.

[22] X. Hu, "Data Mining and Its Applications in Bioinformatics : Techniques and Methods," p. 4577, 2011.

[23] D. Wegener, S. Rossi, F. Buffa, M. Delorenzi, and R. Stefan, "Towards an Environment for Data Mining based Analysis Processes in Bioinformatics & Personalized Medicine," pp. 570–577, 2011.

[24] S. Dixon and X. Yu, "Bioinformatics Data Mining Using Artificial Immune Systems and Neural Networks," pp. 440–445, 2010.

[25] Jacq. N., Blanchet, C., Combct. C., Cornillot, E.. Durct. L., Kurata. K. I., ... & Breton, V. (2004). Grid as a hioinforrnatic tool. Parallel Computing, 30(9), 1093—I 107.

[26] Agrawal, R., Imieliñski, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In ALM SIGMOD Record(Vol. 22, No. 2, pp. 207-2 16). ACM. Atluri, G.. Gupta. R.. Fang. G.. Pandey. G.. Steinhach. M.. & Kumar. V. (2009). Association analysis techniques for biointrniatics problems. Inl)u)u1ormaucs and (‘omputailoital Biology (pp. 1-13). Springer Berlin Heidelberg

[27] Friedman, N., Linial, M., Nachman, 1., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. Journal of computational biology, 7(3-4). 601-620.

[28] Kumar, C., & Mann. M. (2009). Bioinforrnatics analysis of mass spectromeiry-based proleomics data sets. FEBS letters, 583(11), 1703-1712.

[29] S. F. Rosario and K. Thangadurai, "RELIEF : Feature Selection Approach," *Int. J. Innov. Res.*

 Copyright © 2019Authors

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Dev., vol. 4, no. 11, pp. 218–224, 2015.

[30] L. Gao, T. Li, L. Yao, and F. Wen, "Research and application of data mining feature selection based on relief algorithm," *J. Softw.*, vol. 9, no. 2, pp. 515–522, 2014.

Copyright © 2019Authors