IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

A Novel Approach for Classifying Gene Expression Datasets

A. Immaculate Mercy[1*], M. Chidambaram [2]

[1]PG and Research Department of Computer Science, A.V.V. M Sri Pushpam College (Autonomous), Poondi, Thanjavur, India
[2]PG and Research Department of Computer Science, Rajah Serfoji Govt. Arts College (Autonomous), Thanjavur, India

[*]*Corresponding Author: mercybastnj@yahoo.com*

Available online at: www.ijcseonline.org

Accepted: 13/Aug/2018, Published: 31/Aug/2018

***Abstract*—** Classification of Gene expression data is one of the challenging tasks in the domain of Bio-medical recognition. Working on high dimensional data sets always poses complexity on accuracy and on the computational fronts. A Novel approach for classifying the gene expression data has been proposed which paves path for better efficiency and effectiveness measure using an enhanced algorithm for analyzing the sequential patterns by use of a novel algorithm which surpasses the existing methods. This approach provides a better heuristics for working with both supervised and the semi-supervised data. The proposed methodology increases the efficiency by making use of the threshold values which has been used for pruning the data sets which gives rise to a higher confidence on the data sets. The classification thus achieved could help us understand the patterns using the prediction algorithm and then grouping them based on the class labels. This work and the technique that is to be used could serve us in predicting interesting knowledge on the input gene data set. As the data set is of high dimension it throws open the corridors for various analysis on the acquired classes and considerably alleviate the computation cost.

***Keywords*—**Classification, Gene Expression, Supervised, Semi-supervised, pruning

I. INTRODUCTION

Numerous developments in various domains especially health care, science and information technology has felt the need for novel methodologies for working with the data as the dimension of the data has reached its pinnacle. The need for newer methodologies has been on the rise especially when classification of high dimensional structured or unstructured data is considered. The existing works has made use of different classification techniques on both the labeled as well as the unlabelled data for the training process. The semi-supervised extension followed by the self-training scheme with the certainty score increases the size of the labeled instances continually. The inefficient processing of large data sets increases the time complexity and drastically reduces the performance. The classification using support vector machines, KNN, Random Forest suffers from the accuracy as direct feature selection and classification on the combined data sets were best fit only for smaller data sets.

This paper proposes a technique allowing better classification which leads to a better efficiency on the Gene expression data. The process consists of training the data sets from various Gene expressions which are sequence data sets. Data pre-processing is done in the first step which is followed by finding the frequent gene sequence by using the Enhanced algorithm for finding the patterns. This is hen followed by clustering which basically increases the speed of classification which is followed by assigning labels to the gene sequences. Pre-processing is done for the testing data sets. Based on the frequent data the modified classification algorithm is used to find the class labels for the testing Gene Sequence Data sets.

II. RELATED WORK

The study that has been carried for analyzing the various existing methods could be classified based on the approaches and the kind of data sets that has been considered for the work.

1. *Feature Selection and Classification :*

Most of the earlier works has based their works on the common classification methods and Feature selection methods. In the classification methods they have made use of Support Vector machines, K-Nearest Neighbor and Random Forest. The evaluation of classification using Five-Fold cross validation technique has been conducted on the publicly available glioma datasets .The classification was boosted by a small number of genes. The feature selection basically made use of two approaches namely

i. Filter Approach
ii. Wrapper Approach

The filter approach select features as a pre-processing step without considering the classification accuracy. The wrapper approach uses the predictive accuracy of an algorithm to evaluate the possible subset of features that provides highest accuracy. The wrapper methods are computationally expensive and have a higher risk of over-fitting. The Direct feature selection and classification on the combined datasets does not result in good accuracy. The feature selection method focuses on features that are good, rather than strong discriminating groups. Wrapper and embedded are time consuming but yet they could be combined to bring out a hybrid approach.

The other approach that has been cited for the classification process is the use of Ensemble approach. This approach increases not only the classification but also has a strong confidence on the results. These classifiers are less dependent on the peculiarities of a single training set. For the feature selection it makes use of the BAHSIC. The approach that has been implemented led to a fast and adequate system that outperforms the Ensemble system which has been suggested in the earlier works. Only 3 classifiers has been used and tested, but could be extended to multiclass datasets.
The modified AHP allows to process quantitative factors that are ranking outcomes of individual gene selection methods using various statistical methods. An unsupervised learning strategy using the fuzzy e-means clustering was employed. Once again this method also suffers from selection of specific gene selection approach, as it concentrates more on binary classification rather than multiclass problems.

2. *Classification with Microarray data*

Microarray data for clinical medicine plays a significant role owing to its large no. of gene relative to the samples. Determining the subsets of discriminating genes through intelligent algorithms and building a prognosis system leads to challenges. However this approach clearly states that smaller the data set grater the efficiency. Also this method works best on binary data and on multi-class microarray datasets but not on sequential datasets.
The multi-Objective Heuristic algorithm brings out the best gene set by using two attributes:

i. Accuracy maximized
ii. No. of Feature selected minimized

The working of the MODEA is a kind of estimation distribution Algorithm. Though the heuristic approach leads to a well classified model, it suffers in terms of higher computational cost as the algorithm had to go through a lot of iterations which needs to be simplified in the future.

The other method of working with microarray data is based on EM clustering. This work focuses on discretization of genes based on EM clustering. An adaptive selection method is used for exploring the distinct datasets with discrimination power. By monitoring the information gain acquired from the collection of selected features they were able to predict distinction between multiple subclasses without previous knowledge of the subclasses. However this method could be replaced y other clustering algorithms as the size of data grows.

3. *Semi-Supervised Approach :*

The semi-supervised approach of classifying gene expression data has made use of Naïve Hubness Bayesian K-nearest Neighbor. The NHBNN follows a Bayesian approach to assess the probability for an unknown class label. For each labeled training instance, one can estimate the probability of the event that x appears as one of the k-nearest neighbors belonging to a class. Though the occurrences are highly correlated, NHBNN offers improvement over the basic KNN. Self training is one of the most commonly used semi-supervised algorithms. It's basically a wrapper method. Although the Hubness classifier promises accuracy in classification it suffers from analyzing with high dimensional data.

All the above said methods though tried to prove on the fronts of accuracy has somewhere restricted their workings to limited data sets. On the contrary in real life if we need to perform analysis on Gene expression data sets it could not be limited. Therefore a need for a novel approach is needed.

III. METHODOLOGY

The motivation for the novel approach for classification of Gene expression data is twofold. First the need for acquiring a better classification which leads to a knowledge prediction on large structured and unstructured data sets. Two the need for better computation which overcomes the deficiencies of the existing works. This twofold approach has created a niche for the classification of high dimensional data. The data that has been included for the analysis is of the primary form that is sequence data. These sequence datasets are acquired from publicly available gene databases.
The work flow is depicted in the following diagram:

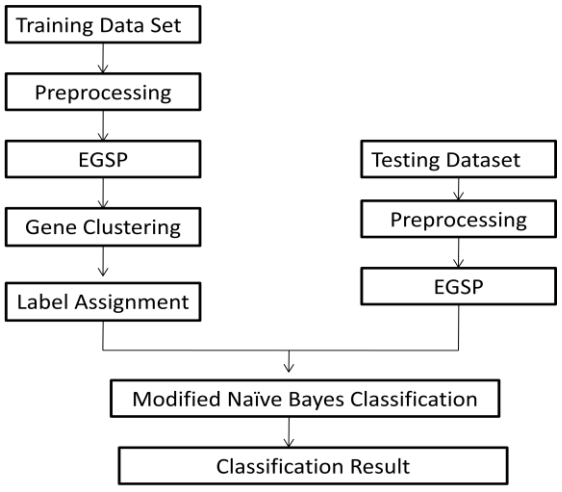

Figure-1

A. Preprocessing the Training and the Testing Dataset

The input for the Training Dataset is the gene expressions which already contains some labels based on the category of Gene selected. These Gene expressions are then passed into a preprocessing phase where only the attributes of interest are considered. The EGSP algorithm plays an important role in finding the frequent patterns, candidate generation and support count.

B. Candidate Gene Sequences Generation

The Candidate sequence generation reads the gene sets. A maximum size for the gene sequences is set after reading through the sequences./ The first step towards candidate generation is the breaking up of individual gene components and finding the distinct components form the sets. A threshold value is set for extracting the candidate sequence. This threshold plays an important part since each and every gene are of varying size based on the structure derived.

The sequential pattern is obtained by performing a weight estimation on the patterns arrived from the previous step. Likewise a class-wise weight is attained. Based on the classes number of sequences is computed .In further support of this process the support and lift is calculated. A threshold is set for pruning the sequences through the maximum and minimum confidence is attained and it is added to the pruned list.

C. Gene Clustering

After the EGSP is run the next step is the clustering of sequences. The phase takes the extracted sequential patterns as input. The cluster heads are selected in such a way that it uses the count of the sequences i.e., the minimum and the maximum count. For all the sequences in the cluster heads a distance is calculated based on the confidence measure. The output of this phase is the clustered sequences.

D. Modified Naïve Bayes Algorithm

After the process for the training and the testing datasets are completed they go into the final phase called classification for which we make use of the Modified Naïve Bayes algorithm. The algorithm takes as input the clustered sequences, where the training and the testing sets are initialized. A count for each class is computed which is followed by the computation of the probabilistic components. A minimal standard deviation rate is set for the testing data set. The probabilistic computation is done for each of the features that have been extracted which falls within the range of difference between the features and the standard deviation rate. Class labels are assigned for each of the classes that have been arrived at.

IV. IMPLICATIONS

The idea of working with these gene sequences poses a great challenge such that not all the gene sequences are labelled. The chances of labelled instances are comparably few when compared to the huge amount of gene expressions are considered. The working of the EGSP and the Modified Naïve Bayes algorithm will tend to be beneficial since all the crucial factors for the classification is considered. The choosing of the threshold value and arriving at the threshold value is a new addition to the already existing algorithms where the traditional algorithms have failed to perform.

As the datasets are surplus the deficiencies that have been encountered in the earlier methodologies will be harnessed in this work. The twofold measure could be handled with ease with the modified algorithms. The accuracy measure also increases the scope of this approach. Proposed work implies to set the discussed phases into an experimental process and the results would be discussed based on the values obtained and statistical approach is brought forth.

V. CONCLUSION and Future Scope

This paper has proposed a study on gene sequences thereby supporting the fact of beneficial approaches for classification for high dimensional structured and unstructured data as well as supervised and unsupervised or the semi-supervised data. As the need for the classification is at the rise, the proposed methodology aims at bringing out a prediction model that could serve the analysis in the domain of biomedical research and sciences in the near future. The analysis would be in perspective of classification performance on various methods, prediction accuracy and accuracy of the specified labels. As the scope for expansion is unlimited for this domain the technique discussed would support the fact of study.

.References

[1] Heba Abusamra. "*A comparative study of featue selection and clasification methods for gene expression data of glioma*", Procedia Science Direct , Elsevier Issue.10.1016/j.procs.2013.10.003. For Conference.

[2] Jia Lv,Qinke Peng,Xiao Chen, Zhi Sun , "*A multi-objective heuristic algorithm for gene expression microaray data classification*", Elsevier, Expert Systems with Applications 59(2016)13-19.

[3] Krisztian Buza, "*Classification of Gene Expression dat: A Hubness-aware semi-supervised approach*", Elsevier, Computer Methods and Programs in Biomedicine 127(2016) 105-113.

[4] Hung-Yi Lin, "*Gene Discretization based on EM clustering and adaptive sequential forward gene selection for molecular classification",*Elsevier, Applied Soft Computing 48(2016) 683-690..

[5] Sara Tarek,Reda Abd Elwahab,Mahmoud Shoman, "*Gene Expression based cancer classification*", Egyptian Informatics Journal 2016.

[6] Devi Arockia Vanitha ,Devaraj D,Venkatesulu, "*Gene Expression Data classification using support Vector Machine and Mutual Information–based Gene selection*", Procedia Computer science 47(2015)13-21.

[7] Konstantina Kourou, Costas Papaloukas,Dimitrois I.Fotiadis, "*Intergration of pathway Knowledge and Dynamic Bayesian Networks for the prediction of Oral Cancer Recurrence*", IEEE 2016.

[8] Thanh Nguyen, Saeid Nahavandi, "*Modified AHP for Gene Selection and Cancer Classification using Type-2 Fuzzy Logic*", IEEE Transtions on Fuzzy Systems, Vol 24 No.2 April 2016.

[9] Jesus Maillo,Sergio Ramirez,Issac Triguero,Fransico Herrera, "*kNN-IS: An interative Spark-based design of the K-nearest Neighbors classifiers for big data*" Knowledge based Systems, Elsevier 000(2016)1-13.

[10]Pradeep K.Sharma, Vaibhav Sharma, Jagrati Nagdiya, *" A Proposed Method for Mining High Quality itemset with Transactional weighted utility using Genetic alogirthm Technique*", IJSRCSE, Vol -5, Issue 1, pp 31-35, 2017.

[11]T. Senthilselvi, R.Parimala, "*Improving Clustering Accuracy using Feature Extraction Method* ", IJSRCSE, Vol-6, Issue-2, pp 15-19, 2108.

Authors Profile

Immaculate Mercy .A is pursuing her Ph.D in Computer Sceince in Bharathidasan University. She holds a Masters Degree in Computer Applications, Masters in Computer Science Engineering and M.Phil in Computer Sceince. She has over 14 years of experience in teaching at the Under Graduate and the Post Graduate level and 8 years of industry experience, which involved in software development, e-content writer. She also has a proven experience in Corporate Training. Her research work focuses on Data Science, Prediction Analysis, Data Mining , Big Data Analytics and Cloud Computing.

Chidamabaram .M pursued his Masters in Computer Science from Bharathidasan r 21 University, M.Phil from Bharathidasan University, M.B.A from Periyar University nad Ph.D from Vinayaka Mission's. in Computer Science. He has over 21 years of experience in teaching at the undergraduate and post Graduate level. He has guided over 25 scholars towards M.Phil degree and 8 scholars towards Ph.D degree. He has published more than 40 research papers in various National and International journals, 8 conference papers.He is currently working as an Assistant professor in Computer sciecne in RSGC, Thanjavur. His areas of Research are Cloud Computing, Grid Computing, Data Mining and Big Data Analytics.

