International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-4, November 2019

Composition of Feature Relevancy Based Biomarker Gene Selection in Gene Expression Dataset

M.Pyingkodi, S.Shanthi

Abstract*— Cancer gene selection plays a prominent work in the area of Bioinformatics. Gene selection methods aim to retain relevant genes and remove redundant genes. This proposed technique deals on gene selection techniques based on information theory. By investigating the information theory based on composition of feature relevancy, we consider that a excellent gene technique method could boost novel classification of the cancer gene data while reducing gene redundancy. Therefore, a modified gene selection technique called Composition of Feature Relevancy (CFR) is carried out. To assess CFR, the experiments are carrying out on five real-world cancer gene expression data sets and three best classifiers (KNN, Support Vector Machine and Random forest). The modified gene selection technique gives best outcome when competing to other recent technique in terms of accuracy and sensitivity in classification.

Keywords : Gene Selection, Composition of Feature Relevancy, Information Theory, Classification, Random forest, Cancer

I. INTRODUCTION

For cancer microarray gene expression dataset investigation, the traditional clustering approach groups the genes over all the conditions or similarly, groups the conditions over all the genes, but in the cellular processes, a subset of genes activeate under a small subgroup of conditions. Further, a individual gene has a chance of linking more than one group and a gene may be entail in more than one biological operations. Therefore it has a higher probability of finding marker genes that are associated with certain tissues or diseases. Hence meta-heuristic bi-clustering approach has been explored as an alternative approach to standard bi-clustering techniques to identify coherent or similar patterns from gene expression datasets.

High dimensional exploration space in microarray gene expression dataset contains an abundant collection of genes with few conditions or samples. All the genes involve in the process are not informative and therefore informative cancer marker genes are needed to be obtained. As the analysis of such dataset is a great challenge while finding the cancer marker genes, analysis of this huge gene dimension diminishes the classifier performance and increases the cost of computation. Without gene selection it is difficult to obtain a satisfactory classification result due to both the over-fitting and curse-of dimensionality problems.

This work, the modified technique provided by the Composition of Feature relevancy (CFR) is examined as a filter approach for the initial feature selection task which gives top 100 genes and GLV performs the second level feature selection method for giving cancer biomarker genes. The gene selection process is performed in two stages. The CFR chooses the most significant features as the first feature selection method. Initially calculating the conditional mutual information and interaction information are performed to find the feature that maximizes the classification accuracy.

Existing cancer gene selection techniques primarily deals on identifying relevant cancer genes. In this paper, feature relevance alone is inadequate for proficient gene selection of high-dimensional cancer gene datasets. Gene selection depends on information theory is employs to identifying a sub-group of the cancer biomarker genes, has sizable application fields in bio-informatics. A novel framework is taken that combines features of relevance analysis and redundancy analysis. The gene correlation-based method is taken genes relevance and genes redundancy analysis for a gene selection and conducts an empirical study of its effectiveness and efficiency comparing with recently arrived methods. Generally group of genes are usually interdependent, where one gene not able to function without another gene. In order to increase the classification accuracy a new filter based cancer gene selection has applied in[1] which taken information theory as fundamental. Moreover, large number of recent approaches endure from mainly two common defect. Feature relevancy is consider without differentiating candidate feature relevancy and taking feature relevancy and few interdependent genes should be miscalculated unwanted features also know as redundant.

In this paper, gene selection method named Composite feature relevancy (CFR) to address these two defects. First, the relevancy of the feature is separated into two classified namely selected feature relevancy and candidate feature relevancy. Next, based on candidate feature relevancy also known as joint mutual information and conditional mutual information, some features that have a redundant data are identified.

Manuscript published on November 30, 2019.
* Correspondence Author
M.Pyingkodi*, Department of Computer Applications, Kongu Engineering College, Erode, Tamilnadu, India. Email: pyingkodikongu@gmail.com
Dr. S.Shanthi, Department of Computer Applications, Kongu Engineering College, Erode, Tamilnadu, India. Email: shanthi.kongumca@gmail.com

Finally, the features that are redundant is minimize the picking the possibility of true repeated genes. The most informative and interdependent features are discovered and true repeated features are removed at the same time.

II. LITERATURE REVIEW

Feature selection techniques are obtainable in literature due to the accessibility of information with hundreds of features leading to data with very huge dimension. Feature selection approaches offers a way of minimizing time taken for computation, increasing the performance of prediction , and a best perceptive of the data in or pattern recognition and machine learning applications. [2] applied ACO for a novel feature selection algorithm known as Advanced Binary ACO algorithm. Gene features are consider as graph nodes to build a graph model and are connected completely each other. Ant colony method is employ to identify nodes while ants must visit all features. The aid of several statistical measures is analyzed as the heuristic method for visibility of the edges in the graph.[3] performed three various feature selection algorithms namely , Backward Elimination Hilbert-Schmidt Independence, Extreme Value Distribution and Entropy gene selection based Singular Value Decomposition, In the ensemble system having a five classifiers like 3-NN method. Each classifier applies its own feature selection parameter to check the variety of the ensemble techniques. The fundamental classifier uses EVD gene selection techniques with automated techniques for giving the number of genes to identifying then the number of genes identified using the method is forty nine genes for Colon cancer dataset.the algorithm returned was two hundred and forty nine genes in Colon cancer dataset and one hundred eight one genes in the Leukemia cancer dataset. Finally integrating the predictions of ensemble members with majority voting. [4] applied Cuckoo search algorithm for gene identification in tumor classification task. In the first process, the top-ten cancer related genes are identified by T-Statistics, signal-to-noise ratio and F-Test. The k-nearest neighbor classifier accuracy is givrn as the fitness function value in Cuckoo search algorithm. The modified algorithm are analyzed and tested with ten various cancer gene expression dataset. For two lung cancer dataset namely Lung Cancer Michigan and Lung Harvard2 cancer the modified method provides 99.25% accuracy in the classification task with reduced size of cancer related genes. For other five datasets, the modified method gives higher than 90% of classification accuracy. The end results conclude that only the cancer related gene selection provides to increases the classifier accuracy. [5] has given doublets, a technique to join cancer gene expression data from gene pairs in the huge dimensions. The selected gene pairs are strong cancer genes as evaluated to single genes from dataset. The genes are communicating to perform a deregulation of the genes in the interaction than genes in individual, may be accountable the critical pathways while deregulating. In the Second phase integrate the concept of doublets with conventional classifiers to work classifiers whose accuracy is greater than that of the original ones. It proved that doublets can be easily combined into the existing classifiers without having to modify the fundamental algorithms, and that using doublets can constantly progress the classification accuracy among various datasets.

From the gene expression data, many genes show inadequate information for performing classification tasks. On the other side, few of them are highly expresses similar behavior , which shows the occurrence of redundant expressed data.[6] provides a technique for the selection of mostly revealing cancer genes with a less redundancy in the dataset. The prognostic accuracy of the identification are assessed using Classification and Regression Trees method. this method allow evaluation of the performance of the identified genes for classify the variable and discovers complicated gene to gene interactions taken [7] introduced a multistage feature selection novel techniques remove unwanted, redundant, noisy and same genes then finds a subset of related cancer genes in the various phase. For removing irrelevant genes SNR algorithm is used in the next process modified algorithm names as support Vector Clustering is used minimize noisy and genes having redundant values. After finding the clusters of gene showing similar behaviour. And then , finding the low positioning genes in the cluster measured as irrelevant genes by SVM-Recursive Feature Elimination method. [8] given an method known as Successive Feature Selection . The SFS techniques of initially finds the partitioning the genes of smaller blocks then identifying significant smaller groups of cancer genes from a subgroup and combines the finding genes with next gene subset (of size) to date the gene subset. The process is continues till all subgroup are combined into one informative sub group .[9] investigate mRMR in the ensemble techniques. The author applied hybrid approach to combine Genetic Algorithm, ReliefF and mRMR design as R-m-GA. In the initial phase, the informative cancer gene set is given to ReliefF. Then, the occurrence of repeated genes is minimize with the help of mRMR, next it assist for the selection of cancer related gene subsets from the candidate set. In the third process, Genetic algorithm gives fitness function and applied to finds the most informative genes. [11] applied rough set and Fuzzy preference based techniques for informative gene indentifying SVM which is based on semi supervised. The performance of the method is evaluate with the Signal noise ratio and Consistency based feature selection methods.[12] Introduced a gene selection techiques in that fixed-point approach is employ for cancer classification using microarray cancer gene expression data set. In the fixed-point method the between-class scatter matrix is employ to calculate the values of leading Eigen vector. The values of Eigen vector has been applied to identify set of genes. [13] Introduced a modified gene selection approach by using regularized linear discriminate analysis which helps to improve performance to find significant cancer involved genes. [14] Recommended an gene selection techniques which employ the online feature selection approach based an online learner is allowed to give a classifier linking a fixed and small number of attributes. [15] introduced a two stage feature selection method. In the stage-1, huge dimensionality of gene set is reduced to hundred of genes. The correlation among the gene is measured and used for classification task. A fast approach is utilizes to manage data with more than three class label.

From Stage 2, the output of the stage 1 is given and finds the optimal smallest subset of genes. [16] employ Ant colony optimization takes global and local pheromone update to identify fuzzy partition that is based on the cancer gene expression value and then createing simpler rule set to identifying the genes set. In order to resolve the continuous and formless expression data of a gene, this work taken artificial bee colony approach to selects the points of membership function. Mutual Information is used for the discovering of informative genes. This hybrid approach is called as hybrid Ant bee algorithm.[17] proposed a efficient novel gene selection approach namely Genetic Algorithm based levy flight for selecting of cancer informative genes from huge microarray gene datasets. This proposed method is robust, accurate, and fast than other recent gene selection methods. Proposed a mixed information gene extraction method combining discrete fourier transform, fisher weight function and principal component analysis. Also, they taken multiple logistic regression analysis combined with the classifier like Bayesian decision to do tumor detection and classification. The experiment outcome conclude that the accuracy 96.80% was given by the classifier for colon tumor gene expression dataset[18].

III PROPOSED WORK

Gene selection techniques aim to keep relevant genes and remove redundant genes. This work concentrates on gene selection approach based on information theory. By comparing the composition of feature relevancy, this approach proves that a best gene selection approach could increasing new classification information while decreasing genes redundancy. To find only relevant genes and remove redundant and nosiy genes, to discover the new classification data contains of gene redundancy and gene relevancy. Huge dimensionality gene expression dataset has to be minimized to perform classification process. To overcome the problem in gene selection methods, a modified composition feature relevancy is employ to perform gene selection method based on combining the various form of information theory and perform gene selection as preprocessing step.

3.1 Significance of Mutual Information

Genes are having importantly different expressions with respect to two various classes are known as deferentially expressed genes. The relevance among the genes are consider as the degree of deferentially expressed genes, which should be computed using mutual information concept. Mutual information for the classes is consider as zero if the value of the gene expression has randomly or uniformly distributed in various classes.

The gene must have large mutual information value if it is exactingly deferentially expressed for various classes applied an improvement to the previous mRMR approaches. Here, the cancer genes are finding based on maximizing the relevance between the gene and the class and then minimizing the redundancies between the feature and the other features. Then, the gene which is most relevant is identified to the outcome set and another solution set is generated comprising of the rest of the genes based on the two selection criteria. Subsequently some genes which satisfy both criteria are selected into the final set. The number of genes in the final set is to be provided by the user.

The mutual information based method is a specific kind of feature selection method which is used to denote the relevance between two random variables. But with this method it is very difficult to compute with high dimensional data. The study focuses on many existing work based on the problems in fixing of with low dimension. For feature selection mutual information is used for high dimensional cases which only support the binary variant. Proposed an efficient mutual information based gene selection algorithm in which genes are selected based on the approximate measure of mutual information calculated between the class and the selected genes. An effective pruning strategy is introduced in the selection to improve the efficiency of this algorithm [28].

3.2 Information Theory

The CFR method is working on information theory for feature selection performance. Huge volumes of filter techniques consider on information theory are derived to maximize the relevancy among classes and features while decreasing the redundant features among them. Information theory concept is a significant parameter that is widely employ in filter approaches. It involves conditional mutual information, mutual information, interaction information and joint mutual information.

Consider three variables X, Y and Z are as discrete random. The entropy and conditional entropy are used to find the mutual information. Entropy gives the information like uncertainty of a variable in random and conditional entropy provides the amount of uncertainty remain when introducing another variable. Moreover, the uncertainty reduced is when another variable is introduced and taken as mutual information, it also gives the amount of information that both variables are sharing.

The conditional mutual information is consider as another significant parameter in the information theory The Conditional mutual information taken X,Y and Z variables to calculate information theory. the amount of information among X and Y is calculated while Z is commenced. Likewise, Joint mutual information among the variable X , Y and the variable Z. The information of Joint Mutual information indicates the variable X,Y and Z is used assess feature information redundancy. Interaction information are applied to evaluate redundancy occurs in the feature set in the task of gene selection Interaction information is to zero, negative or positive . Then it shows positive information when the variables are Y and Z is to be provides the similar information; nest it is gives negative when Z and Y provides more information than it give individual information of interaction information ; it shows zero when Y and Z are independent each other. In other sources, interaction information of the variable are derived by the conflicting sign in Equation (1); however, this is for mathematical purpose and does not modify the final conclusion.

$$J(X_k) = \sum_{X_j \varepsilon S} \left\{ I\left(X_k; \frac{y}{x_j}\right) - I\left(x_k; y; x_j\right) \right\} \quad (1)$$

Similar to JMI and mRMR in proposed method, the redundancy term weight is computed and it is negatively related with the relationship of the gene feature subgroup. It is useful to decrease the impact of the repeated term that is redundancy term while maximizes the selected genes. The CFR approaches and the analysis of recent techniques together employ identical effective greedy searching method, knows, Sequential Forward Search. The gene subgroup selection S begins as an null set, next, finds one informative gene consider on the feature selection approach each and every time till the amount of identified features as genes is higher than the prescribed given threshold value t

3.3 Composition of Feature Relevancy Pseudo code

The procedure of CFR method are illustrate as follows:

1. Initialization Set O? "original Feature set of n features", E ? "empty set"
2. Calculate mutual information between the class with each candidate feature for each feature xt ? O, Calculate I(x_t ;Y).
3. Select the first feature find the feature that maximizes I(x_t ;Y), O? O\{x_t}; E? {x_t}.
4. Greedy Selection repeat until |E| = t
   calculate conditional mutual information and interaction information for all p airs of variables X_i and X_j such that Xt ? 0, X_i ? E, calculate I(X_t; Y|X_i) and I(X_t; Y; X_i).
   select the next feature choose the feature x_t that maximizes formula
   $J(X_k) \quad \sum_{X_j \varepsilon S} \left\{ I\left(X_k; \frac{Y}{X_j}\right) \quad I(x_k; y; x_j) \right\}$
3. The output is the set E that includes the selected features

The algorithm, finds the most appropriate relevant genes as the gene selection phase. Then, computed information of both interaction information and conditional mutual information are computed to find the feature that increases the criterion of CFR. The procedure is ended when |E| = t.

IV EXPERIMENTAL RESULT ANALYSIS

Before conducting the experiments, data pre-processing is a vital process since data is in noisy manner. The data sets must be normalized before pre-processing step taken. The min-max normalization is taken as initial pre-processing process which aid to reduce the expression measurements variation is used to normalize the data sets. The gene expression values computed and scaled to the lowest value for each gene becomes zero and assign one for the highest value. In this work, various cancer microarray gene expression datasets are applied to assess the performance of CFR algorithm. The improvements of CFR method and other popular algorithms for cancer related gene selection task using five microarray cancerous datasets.

4.1 Complexity Analysis

The K represents the reduced number of features or gene; M denotes the number of instances in the dataset, and N defines the number of genes or features. The time complexity of the three information theory are conditional mutual information, mutual information and joint mutual information is given as O(M) however all records need to be computed for the probability estimation. The time complexity of O(kMN) is obtained through this proposed method. Hence the time complexity of proposed techniques is acceptable.

(5.6)

Table 1. Experimental result of 50 independent runs using CFR on Colon

Algorithm	No. of Genes	Accuracy	Sensitivity	Runtime (in Seconds)
COA	8	99.46	96.06	734.97
GA	9	99.81	98.95	849.77
GLV	10	99.77	99.45	98.3
COA- HS	8	99.16	97.51	696.49
CFR	9	99.34	97.36	614.06

Table 2. Comparison of CFR method with various methods on different Cancer Datasets in Random Forest Classifier

Algorithm	Colon	Leukemia	Lung	Lymphoma
mRmR_PSO (Javad 2013)	93.55 (78)	95.83 (53)	94.79 (65)	96.96 (82)
mRmR_GA (Akadi 2011)	95.6 1(83)	93.05 (51)	95.83 (62)	96.96 (5)
mRmR_ABC (Alshamlan et al. 2015)	96.77 (15)	100 (14)	100 (8)	100 (5)
GBC(Hala 2015	98.38 (10)	100 (4)	100 (4)	100 (4)
CFR	99.22 (5)	100 (4)	100 (4)	100 (4)

From the Table 2 shows the accuracy of classifier, the recent gene Selection methods under evaluation when integrated with random forest as a classifier for four microarray dataset. The numbers inside the parentheses consider the numbers of selected genes involved in classifier.

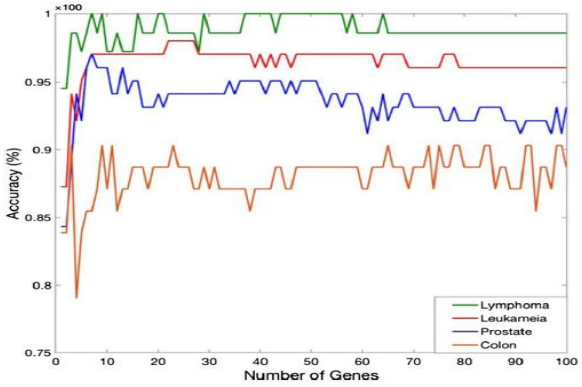

Figure 1 Comparison of Reduced no. of Genes with Accuracy on CFR

Figure1 shows that the CFR select top 9 genes and produces the accuracy of 99.34 and sensitivity of 97.36 for colon dataset which is better than the other feature selection algorithms. Moreover, for a few instances, the classification accuracy was minimized as the number of genes increased. For example, in the colon cancer gene expression dataset, the accuracy given by the classification algorithm for the eight genes shows 97%, but the performance of the accuracy was minimized while the number of genes increase, obtaining accuracy value of 91 to 93% when 80 to 100 genes are taken.

V. CONCLUSION

CFR based gene selection technique is employ for identifying a subset of genes. The mutual information is primarily considered for the elimination of irrelevant and redundant genes from the huge dimension cancer datasets in the first phase. Moreover this method concentrates on relevancy of the selected subset of genes using joint mutual information and condition mutual information for calculating the relevance among the selected genes. This method combines both filter and information theory approach to enhance the performance of the classifications task. The selected subset genes are evaluated by computing the classification accuracy of three algorithms namely KNN, Support vector machine and Random Forest

REFERENCES

1. Xin Sun,Yanheng, Jin, 2012, 'Feature Evaluation and Selection with Cooperative Game Theory", Journal of Pattern Recognition, vol. 45, no. 8, pp. 2992-3002.
2. Shima Kashef, N & Hossein Nezamabadi-pour 2015, 'An advanced ACO algorithm for feature subset selection', Neurocomputing, vol. 147, pp. 271–279.
3. Sara Tarek, Reda Abd Elwahab & Mahmoud Shoman 2017, 'Gene expression based cancer classification', Egyptian Informatics Journal, vol. 18, pp. 151–159.
4. Gunavathi,premalatha, "cuckoo search optimization for feature selection in Cancer Classification: A new Approach", International journal of Data mining and Bioinformatics, Vol. 13, no. 3, oct 2015, pp. 248-265.
5. Chopra, P, Lee, J, Kang, J & Lee, S 2010, 'Improving cancer classification accuracy using gene pairs', PLoS ONE, vol. 5, no. 12, doi:10.1371/journal.pone.0014305.
6. Arevalillo, JM & Navarro, H 2013, 'Exploring correlations in gene expression microarray data for maximum predictive-minimum redundancy biomarker selection and classification,' Computers in Biology and Medicine, vol. 43, no. 10, pp. 1437-1443.
7. Du, W, Sun, Y, Wang, Y, Cao, Z, Zhang, C & Liang, Y 2013, 'A novel multi-stage feature selection method for microarray expression data analysis', International Journal of Data Mining and Bioinformatics, vol. 7, no. 1, pp. 58-77.
8. Sharma, A, Imoto, S & Miyano, S 2012, 'A top-r feature selection algorithm for microarray gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinformatics (TCBB), vol. 9, no. 3, pp. 754–764
9. Shreem, S, Abdullah, S, Nazri, M & Alzaqebah, M 2012, 'Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection', J. Theor. Appl. Inf.Technol., vol. 46, no. 2, pp. 1034–1039.
10. Sharma, A, Imoto, S, Miyano, S & Sharma, V 2012b, 'Null space based feature selection method for gene expression data', International Journal of Machine Learning and Cybernetics, vol. 3, no. 4, pp. 269-276.
11. Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D. & Maulik, U. Recursive Memetic algorithm for gene selection in microarray data. Expert Systems with Applications 116, 172–185 (2019).
12. Sharma, A, Imoto, S & Miyano, S 2012a, 'A top-r feature selection algorithm for microarray gene expression data', IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 3, pp. 754-764.
13. Sharma, A, Imoto, S, Miyano, S & Sharma, V 2012b, 'Null space based feature selection method for gene expression data', International Journal of Machine Learning and Cybernetics, vol. 3, no. 4, pp. 269-276.
14. Wang, J, Zhao, P, Hoi, SCH & Jin, R 2014, 'Online feature selection and its applications', IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 3, pp. 698-710.
15. Chakraborty, G & Chakraborty, B 2013, 'Multi-objective optimization using Pareto GA for gene-selection from microarray data for disease classification', Proceedings of IEEE international conference on systems, man, and cybernetics (SMC), pp. 2629-2634.
16. Ganesh Kumar, P, Rani, C, Devaraj, D & Victoire, TAA 2014, 'Hybrid ant bee algorithm for fuzzy expert system based sample classification', IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 11, no. 2, pp. 347-360.
17. Majid Mohammadi, Hossein Sharif Noghabi, Ghosheh Abed Hodtani, Habib Rajabi Mashhadi, 'Robust and stable gene selection via Maximum-Minimum Correntropy Criterion', Genomics, vol. 107, pp. 83-87.
18. Jun Yao , Yuchun Zhang & Pingbo Hao 2014, 'Research on classification and detection of colon cancer's gene expression profiles', Journal of Chemical and Pharmaceutical Research, vol. 6, no. 7, pp. 2792-2800.

AUTHORS PROFILE

M.Pyingkodi is an Assistant Professor in the Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India. She received her Bachelor's degree in Computer Science at 2003 and a Master's degree in Computer Applications from Bharathiar University at 2006. Her areas of specializations are Data Mining in Bioinformatics, having teaching experience around eleven years. She is pursuing her Ph.D in Computer Science at Anna University, Chennai.

Dr. S. Shanthi received her PhD degree in Computer Science and Engineering at Anna University, Chennai, India in 2015. She is presently working as an Assistant Professor (SLG) in the Department of Computer Applications, Kongu Engineering College, Tamil Nadu, and India. Her area of interest includes, Data Mining, Image Processing, Pattern Recognition, Big data analytics, Health care Informatics and Soft Computing.

D.Hemalatha is an Assistant Professor in the Department of Computer Technology, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India. She received her Bachelor's degree in Computer Science at 2005 and a Master's degree in Computer Applications from Bharathiar University at 2008. Her areas of specializations are Cloud Computing,Data Mining in Bioinformatics, having teaching experience around eleven years..