JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 6, ISSUE 05, 2019

Interdependent Gene Selection for Disease Diagnosis

Smruti Smaraki Sarangi[1] & Shikha Tiwari[2]

Department of Computer Science and Engineering, Kalinga University, Naya Raipur, Chhattisgarh
*Email:smruti.sarangi@kalingauniversity.ac.in

ABSTRACT

Studying of gene expression technology is one of the known ways for proper diagnosis of diseases. But the problem here is the enormity of data which makes it very difficult for biologists to study in a short time. This is where the picture of Gene Selection comes. Gene selection gives a few very informative genes, which are responsible for a particular disease, to the Machine Learning Algorithm and computation is done in a comparatively very less time.

A number of studies have been done in this respect and several methodologies are proposed, and tested on datasets, which have given good prediction accuracy. Notable among these methods are Dynamic Relevance (Sun et.al., 2013), Pathway based approach which requires prior biological knowledge, some hybrid methods, and other information theory based approaches like information gain, distance measure and some metric based measures like m-RMR, T-Statistic measure (Khosgoftaar et.al., 2013). Some of these methods will be compared on different classifiers and the resulting subsets aggregated to find new subset in my model.

INTRODUCTION

Biologists and doctors perform diagnosis of disease in two ways: (a) by proper medical diagnosis. (b) by studying the gene expression data set.

The problem with medical diagnosis is the time taken for different tests to be carried out. This time can't be reduced since it is the machinery and physical work. This problem can be addressed by diagnosing the micro array data set. But this approach has a initial bigger time consuming aspect of its own where biologists have to study the expressions of thousands of Genes and samples to come to a conclusion on diseased and non-diseased patients for future use. But once the initial step is completed, Doctors can just view the Gene data set and say about a patient without going through time consuming medical tests. This problem in the initial stage is called the problem of plenty or ENORMITY of DATA. This problem has a answer in data mining where, based on medical studies that only a few genes are responsible for a particular disease, Feature selection technique, known as GENE SELECTION in case of micro array data set, is used to suggest a few percentage of genes to a machine learning algorithm.Selection of relevant genes for sample classification is a common task in most gene expression studies, where researchers try to identify the smallest possible set of genes that can still achieve good predictive performance (for instance, for future use with diagnostic purposes in clinical practice). Selection of relevant genes saves time in computation with respect to medical diagnosis.Genes normally work in group.

That is some genes are effective when they work as a group. But independently they do not contribute to the cause of a disease. These genes form a gene group. Most of the genes in a gene group have some similar features. These genes are called interdependent genes, i.e. they depend on one anotherfor their functionality. In other words the proteins produced by these genes interact among themselves to perform some functions. So if we use a gene selection algorithm to find a subset of genes, then it will select one gene and discard the rest which are highly correlated to the selected gene.

This disadvantage is fatal to disease diagnosis and classification. The main reason is the interdependent gens. In this work, we discuss about certain approaches where the above problem is tackled by some methods different from the traditional approaches. The beauty of these methods is that it not only selects the most relevant genes and eliminates redundant genes, but this elimination process also takes into picture the interdependence feature, which is the background of the method, where the foremost aim is to retain the intrinsic interdependent gene groups.

ISSN- 2394-5125 VOL 6, ISSUE 05, 2019

Then the selected subset can be applied on any classifier and find a classifier which gives the best prediction of samples based on the selected subsets.

LITERATURE SURVEY

The authors here proposed a new filter technique for gene selection. This is an information theory based measure. (Sunet al., 2013)

At the beginning, the desired selected subset GS is initialized as an empty set, and the dynamic relevance value DR(g) of each gene is initialized to the mutual information I (g, class).In each iteration, the first gene with the highest priority is chosen into the gene set GS. Then the algorithm calculates the correlation ratio of each remaining gene with the newly selected one, and updates dynamic relevance values for every gene. The selection procedure will be terminated if the number of selected genes is larger than the user-specified threshold.

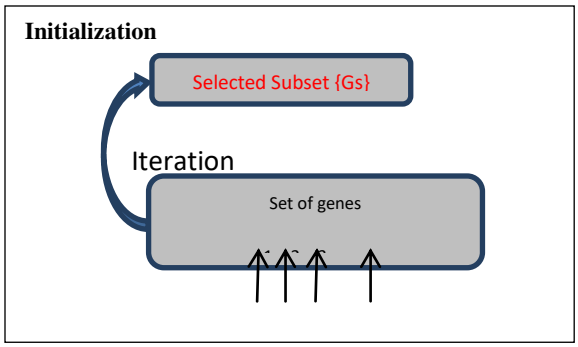

Fig 1: Initialization Step

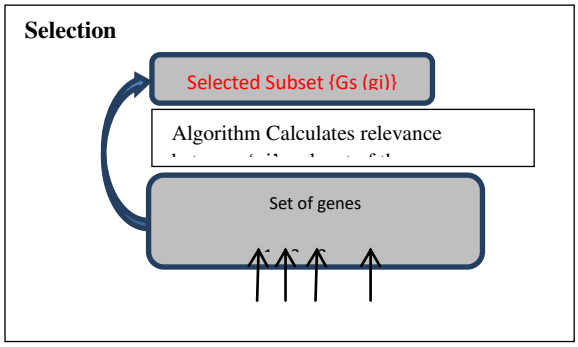

Fig 2: Continuous selection step

The relevance here is called dynamic relevance since it is updated at each iteration based on the selected genes into the subset. A threshold has to be provided on the selected-set as for number of genes to be selected or asa minimum value of the dynamic relevance.
Dynamic relevance approach was applied on Breast cancer data set, Central nervous system data set, Lung cancer data set, Prostate cancer data set, Gastric cancer data set, Childhood ALL data set. Better prediction performance was observed than ReliefF, SAM, m-RMR and IG when tested on SVM and
K-NN.

Khoshgoftaar have compared a present ensemble technique with another two proposed ensemble techniquesin their paper. (Diitmanet al., 2012) The focus of creating an ensemble feature selection technique is how to approach the concept of ensemble. Ensemble feature selection is a subset of feature selection techniques which applies feature selection algorithms multiple times and combines the results into one decision.The idea for ensemblefeature

ISSN- 2394-5125 VOL 6, ISSUE 05, 2019

selection is derived from ensemble learning methods wherein different classifiers are applied to a data set and their results are aggregated. We have decided upon comparing three approaches of ensemble: data diversity, functional diversity, and a hybrid approach which combines data and functional diversity.

Data Diversity

Data diversity, as its name suggests, achieves its diversity through the use of different sets of data. The process for data diversity occurs in three steps. The first step involves creating the different data sets in order to achieve thedesired diversity. This can be achieved through the use of different compileddata sets which use the same set of features or, more commonly used, thecreation of multiple sets of sampled data derived from the original data set.The next step is to apply the same feature selection technique on each ofthese new data sets. Lastly, we aggregate the results from each of the datasets and end with a single feature subset for use in subsequent analysis. Allrecent works discussing the use of ensemble feature selection techniques usedata diversity when creating ensemble techniques.

Functional Diversity

Functional diversity uses a completely different methodology than that ofdata diversity. In functional diversity, the same data set is used throughoutthe entire process. The process of functional diversity takes the original data set and applies a number of different feature selection techniques to create aranked list for each technique. After all of the chosen techniques have been performed, the results are aggregated into a single feature ranking. To ourknowledge, there has been no study which uses functional diversity within the domain of bio-informatics.

Hybrid Approach

The two previous techniques take vastly different paths in order to achievetheir diversity. Both paths have their benefits and detriments when being implemented. However, using one of these approaches does not preclude theimplementation of the other approach. As with functional diversity, there has been no studywhich uses hybrid methods within the field of bio-informatics. 26 different data sets were used and 5-NN found to be best learner. Comparing these two approaches to the existing data diversity approach, they perform better in terms of classification.

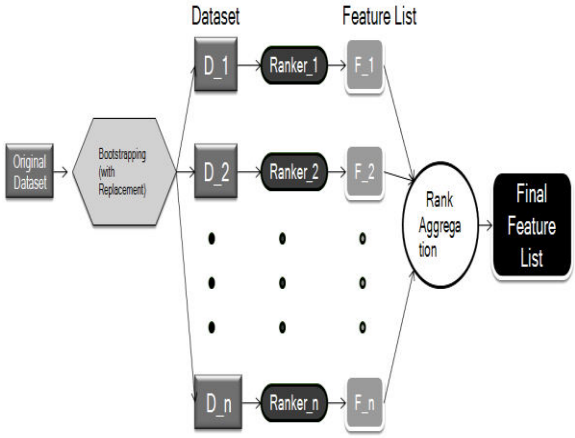

Fig 3: Hybrid Approach

Ibrahim have proposed a pathway-based approach for the problem. (Ibrahim et.al, 2011). In the proposed method it requires to have some basic biological knowledge to understand what is fold change and pathways.

First a fold change value is or attribute is added to all the genes in the dataset. Then it is mapped onto the pathway imported into mat lab from the Kyoto Encyclopedia of Genes and Genomes database (KEGG). Each pathwayhas certain genes which interact within themselves. These pathways are thenranked according to their Z-Score. The z-score is a standard statistical test. It has been used as a measure of significant pathway perturbation after superimposing micro array expressiondata. The z-score value is calculated simply by subtracting the

ISSN- 2394-5125 VOL 6, ISSUE 05, 2019

expectednumber of genes meeting the criterion from the observed number, and thendividing by the standard deviation of the observed number of genes as follows:

$$Z - Score = \frac{r - n\frac{R}{N}}{\sqrt{n\frac{R}{N}(1-\frac{R}{N})(1-\frac{n-1}{N-1})}}$$

Then the genes of the top ranked pathways are ranked according to their fold change.This is the selected subset.It is clear that using combined pathways provides higher accuracy compared to a single pathway.

$$FG_i = \begin{Bmatrix} abs(F_{c_i}); (F_{c_i} \geq h, p < p_i) \\ 1; otherwise \end{Bmatrix}$$

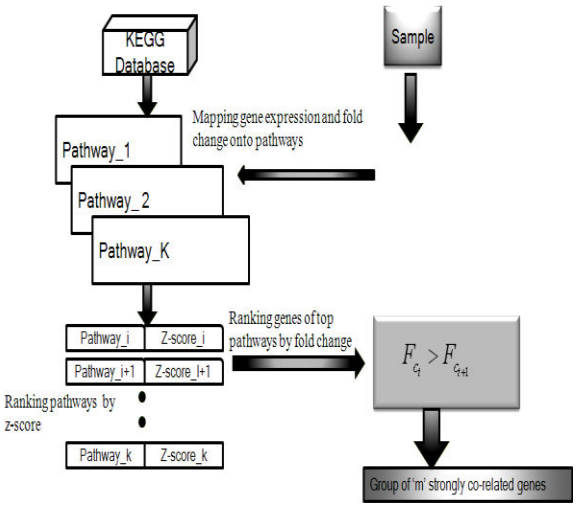

Fig 4: KEGG pathway

Ganesh Kumar did gene selection in two steps. (Kumaret.al, 2011). First it applies mutual information technique.Mutual information is a symmetrical measure. That is, the amount of information gained about Y after observing X is equal to the amount of information gained about X after observing Y. For the gene selection problem under consideration, X corresponds to the set of input genes and Y corresponds to the class label.
After that augmented variance ratio is calculated as:

$$AVR = \frac{S_b}{\frac{1}{M}\sum_{m=1}^{M}\frac{s_m}{min_{m \neq n}|\mu_m - \mu_n|}}$$

where M is the number of classes, $S_m$is within class variance for the m^{Th}class,Sb is the between class variance, μ_m is the mean for the m^{Th} class and thus is
μ_n for the nth class. Thus the AVR imposes a penalty on features which mayhave small intra-class variance but which have close interclass mean values.
Multilayer Feed Forward neural network learned by back propagation algorithm used for classification.This Proposed approach selects highly informative genes.

Rahidehentitled Cancer classification using clustering based gene selection and artificial neural networkused a statistical measure for gene selection same as the signal to noise but given a different name of f-score in their paper. (Rahideh& Shaheed, 2011).

ISSN- 2394-5125 VOL 6, ISSUE 05, 2019

Khoshgoftaar entitled Feature List Aggregation Approaches for Ensemble GeneSelection on Patient Response Datasetsin their paper. (Khosgoftaaret.al., 2013).The objective of this paper was how to aggregate the resulted gene subsets from different gene selection algorithms. They used three different feature selection algorithms in this research: Information Gain (IG), Area under the Receiver operating characteristics Curve (ROC), and Signal-to-Noise (S2N). Each of these techniques are filter-based feature ranking techniques.

IG

Information Gain (IG) is one of the simplest and fastest feature ranking techniques, and is thus popular in bio informatics where high dimensionalitymakes some of the more complex techniques infeasible. IG determines the significance of a feature based on the amount by which the entropy of theclass decreases when considering that feature.

At first the expected information needed to classify a tuple in D is given by:

$$info(D) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

How much more information would we still need in order to arrive at an exact classification? This amount is measured by:

$$info_a(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} info(D_j)$$

Gain is calculated by:

$$Gain_A = info(D) - info_A(D)$$

ROC

Receiver Operating Characteristic, or ROC, curves are a graph of the true positive(sensitivity) rate on the y-axis versus the false positive rate on the x-axis. In the context of feature selection, this is found by considering a datasetwhich contains only the class feature and the feature being evaluated. Valuesabove a certain threshold are considered positive instances, while those below the threshold are negative instances. The reverse, with values above thethreshold being negative and those below being positive is also tested andthe direction with the best result is used. The ROC curve is graphed as thisthreshold varies, and the curve itself represents the trade-off between the rate of detection and the rate of false alarms. The area under this curve (AUC) is thus used as a single-value metric for the importance of the feature.

Signal to noise (S2N)

S2N is less often used in the context of feature selection.The signal-to-noiseratio, or S2N, as it relates to classification or feature selection, represents howwell a feature separates two classes.

The equation for signal to noise is:

$$S2N = \frac{\mu_p - \mu_n}{\sigma_p + \sigma_n}$$

They had applied these techniques on 15 data sets to generate their respective subsets.Then they used some aggregation technique which is beyond my scope.

The researcher Sakarand his team gave a good overview of the minimum Redundancy Maximum Relevance (m-RMR) approach. This strategy focuses on the fact that individually good genes donot necessarily give good classification accuracy as a group. Thus to improve the joint classification accuracy the redundancy among them should be reduced. This approach is based onthe information theory. They have proposed an improved approach on m-RMR called as KernelCanonical Correlation Analysis m-RMR (KCCAm-RMR) where during the calculation of mutualinformation, instead of the feature 'Xi' the correlated function 'fiu (Xi)' is used. Where 'fiu (Xi)'denotes the various relations of 'Xi' with target class 'T'. Thus a filtering out is performed toremove all the irrelevant relations.

ISSN- 2394-5125 VOL 6, ISSUE 05, 2019

Mandal and Mukhopadhya (2015)have proposed an improvement to the existing m-RMR approach. In their experimental evaluation on feature selection they have selectedthe features based on maximizing the relevance between the feature and the class and then minimizing the redundancies between the feature and the other features. Then the feature which ismost relevant is selected to the output set. Then another solution set is created comprising ofthe rest of the features based on the two selection criteria. Subsequently some features whichsatisfy both criteria are selected into the final set. The number of features in the final set is tobe provided by the user.

Yibing Chen did the feature selection in two phases. (Chen et al., 2011). In the first phase they usedBhattacharyya Distance to separate the non-informative genes to create a smaller set of informative genes. In the second phase they used kernel distance as a strategy to measure the classes'parability which is a way of Floating Sequential Search Method (FSSM). They have applied this on a colon cancer data-set with SVM as the classifier and achieved worthy results.

The researcherproposed a new distance measure named as "Max-range Distance" to compute similarity between two genes..In this approach normalization is first done on the distance between two genes. The normalizing factor is different for the different experimentswhich give the data-set although it is similar for all the genes in the data-set. They took thenormalizing factor as the 'linear dynamic range' of the 'Photo Multiplier Tube' which is used to scan the 'Fluorescence intensities' of that experiment.

Piramuthuused the Hausdorff Distance for feature selection which is the measure of similarity between two features in metric space. The distance is calculated between the features of two different classes. Then a decision tree is constructed by taking the distances in ascending order. The tree stops on a stopping criterion and then the features are sent to theselected subset. The quality of the tree is evaluated by classifying unseen examples.

Hsu et al, (2011)in their study used two different approaches to feature selection approaches i.e. Euclidian distance and Pearson Correlation Coefficient both of whichare statistical measures.

In using 'ID3' as a decision tree algorithm, it uses an attribute or feature selection measure called as information gain. In the year 1948, Claude Shannon developed the idea of information entropy which is the measure of uncertainty in a message and then further digging into it to lay thebase knowledge for information theory where the 'information content' of a message is calculatedmathematically. Using it in terms of a micro-array data-set and in decision trees, we know thatdecision tree is a tree where each node is actually a mini data-set where based on an attribute
the data-set can be divided or classified. Now let we have a node 'N' where we hold the data-setor a part of it referred to as 'D'. Now here we calculate the information gain of all the attributesand the attribute which will have the highest information gain value is selected as the attributefor dividing the data-set into a number of partitions.This attribute is chosen as the splittingattribute.

SUMMARY

Filter method is mostly preferred since here the selection process is done only once and then can be applied to a number of classifiers. Also this method is faster than other methods.This meets the most important objective of gene selection which is faster computation.

REFERENCES

1. Ibrahim, M. A. H., Jassim, S., Cawthorne, M. A., &Langlands, K. (2011, June). Pathway-based gene selection for disease classification. In *International Conference on Information Society (i-Society 2011)* (pp. 360-365). IEEE.
2. Chen, Y., Zhang, L., Li, J., & Shi, Y. (2011, August). Domain driven two-phase feature selection method based on Bhattacharyya distance and kernel distance measurements. In *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* (Vol. 3, pp. 217-220). IEEE.

ISSN- 2394-5125 VOL 6, ISSUE 05, 2019

3. Kumar, P. G., Rathinaraja, J., & Victoire, T. A. A. (2011) "A combined MI-AVR approach for informative gene selection".*Proceedings of second international conference on sustainable energy and intelligent system(SEISCON 2011)* , July 20-22, pp.870-875, 2011
4. Ganivada, A., Ray, S. S., & Pal, S. K. (2013). Fuzzy rough sets, and a granular neural network for unsupervised feature selection. *Neural Networks*, *48*, 91-108.
5. Guyon, I., &Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.
6. Hsu, H. H., Hsieh, C. W., & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, *38*(7), 8144-8150.
7. Dittman, D. J., Khoshgoftaar, T. M., Wald, R., & Napolitano, A. (2012, December). Comparing two new gene selection ensemble approaches with the commonly-used approach. In *2012 11th International Conference on Machine Learning and Applications* (Vol. 2, pp. 184-191). IEEE.
8. Khoshgoftaar, T. M., Wald, R., Dittman, D. J., & Napolitano, A. (2013, August). Feature list aggregation approaches for ensemble gene selection on patient response datasets. In *2013 IEEE 14th International Conference on Information Reuse & Integration (IRI)* (pp. 317-324). IEEE.
9. Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), 273-324.
10. Mandal, M., & Mukhopadhyay, A. (2015). A Comparative Study Among Various Statistical Tests Using Microarray Gene Expression Data. *Current Bioinformatics*,*10*(4), 377-392.
11. Rahideh, A., & Shaheed, M. H. (2011, December). Cancer classification using clustering based gene selection and artificial neural networks. In *The 2nd International Conference on Control, Instrumentation and Automation* (pp. 1175-1180). IEEE.
12. Sakar, C. O., Kursun, O., &Gurgen, F. (2012). A feature selection method based on kernel canonical correlation analysis and the minimum Redundancy–Maximum Relevance filter method. *Expert Systems with Applications*, *39*(3), 3432-3437.
13. Sun, X., Liu, Y., Wei, D., Xu, M., Chen, H., & Han, J. (2013). Selection of interdependent genes via dynamic relevance analysis for cancer diagnosis. *Journal of biomedical informatics*, *46*(2), 252-258.
14. Wickboldt, A. K., &Piramuthu, S. (2012). Patient safety through RFID: Vulnerabilities in recently proposed grouping protocols. *Journal of medical systems*, *36*(2), 431-435.