MULTI FILTER ENSEMBLE METHOD FOR CANCER PROGNOSIS AND DIAGNOSIS

Bibhuprsad Sahu
Department of CSE
Gandhi institute For Technology,
Bhubaneswar, Odisha, India

Abstract— The nature of the gene expression profiles are high dimension, very small sample size, continuous types so it is really a challenged task to achieve good classification accuracy from the tumor samples. The main aim of feature selection is to find out most relevant features, which may increase the computational speed and accuracy. We thus proposed a multi filter ensemble based hybrid gene selection method. Here we have used four filter methods such as Information Gain (IG), Gain ratio (GR), Relief, Correlation to filter the irrelevant and redundant genes. By the help of computationally efficient filters candidate features are selected. The top N genes with highest rank of individual subset are integrated to produce a new dataset. Then SVM attribute evaluator is applied for attribute evaluation purpose. Finally LIBSVM classifier is used to detect the best feature subset. This experimental result proves the proposed method is quite efficient then other gene selection methods and it provide a high accuracy under some characteristics genes.

Keywords— IG, GR, CORRELATION, ReliefF, LIBSVM

I. INTRODUCTION

Due to availability of large scale genes expression profiles, for disease diagnosis DNA chips are used to represent expression level of millions of gene in the a single sample for single experiment. To achieve a great significant level of diagnosis for cancer treatment, molecular level accurate classification is needed. The size of tumor gene expression data is high dimensional and small size in nature[1-4]. As the gene expression profiles consists of high redundancy and noise data, so choosing a better filter approach to minimize the computational time with good classification accuracy[5-6]. Each and every sample consists million of genes but from these few genes are really responsible for sample classification. To train an effective classification model one should have a clear idea how to filter the group of related genes form the huge dataset. That is known as coarse of dimensionality problem[7].

While implementing subset feature selection it selects the all attributes of the gene expression profile data [8] and from this it identify the strong attributes to identify the disease. Gene selection methods are generally two types such as filter and wrapper [9, 10]. For gene selection the filter methods uses feature ranking approaches. The filter methods are of two different types such as univariate and multivariate. The Univariate methods are based on single criterion principle nothing to do with the feature selection process such as SNR [11], TS [12], FT [13], PC [14]. By the help of filter methods we have consider the top ranked genes as biomarker genes. In this research we have used Information Gain (IG) [15], Gain ratio (GR), ReliefF [16], Correlation as filter methods.

Filter method for feature selection depends on the individual vector data and final subset selection is independent of classifiers. According to the basic principle of a classifier the wrapper methods detects the biomarker gene datasets. Selection of feature subset done by different machine learning algorithm [17-19], the next evaluation is done by the classifiers[20]. The computational speed of filter methods are fast as compared to wrapper but the relation between genes are not established by the filter [21]. To achieve good accuracy various hybrid methods has been proposed by different researchers[ 22-26].

In this research study, the overall proposed model is divided into 3 stages such as filter gene selection stage, combinational stage, classification stage. In filter gene selection stage is used to identify and remove redundant and irrelevant genes. IG, GR, ReliefF, Correlation are chosen as ranker algorithm. These four resulted top ranked features are combined together for further fine tuning purpose. This is designed at combinational stage. Later stage using SVM filtration and reduction of redundant genes performed. At last LIBSVM classifier is used to obtain the informative genes through classification.
II. HYBRID FEATURE SELECTION

The purposed model is basically divided into 4 stages such as filter genes selection stage, Combination stage, SVM attribute selection stage, Classification stage. The discussion of individual stage is done below.

A. Original dataset –

For this research experimental study, we have considered five bench marked microarray datasets. These five datasets are having two class such as normal and cancerous. Datasets chosen are leukemia, prostate cancer, Colon cancer, DLBCL, lung cancer. We have used AML and ALL (Leukemia) type where as the DLBCL dataset of two sub groups i.e. germinal center B-cell like subgroup, activated B-cell subgroup. The details of gene expression dataset are mentioned below.

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>No of Genes</th>
<th>Sample count</th>
<th>Positive Samples</th>
<th>Negative Samples</th>
<th>No of Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>7129</td>
<td>72</td>
<td>25</td>
<td>47</td>
<td>2</td>
</tr>
<tr>
<td>DLBCL</td>
<td>4026</td>
<td>47</td>
<td>24</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>Colon</td>
<td>2000</td>
<td>62</td>
<td>40</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Lung</td>
<td>7129</td>
<td>96</td>
<td>86</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>Prostate</td>
<td>12600</td>
<td>102</td>
<td>52</td>
<td>50</td>
<td>2</td>
</tr>
</tbody>
</table>

B. Filter gene selection stage-

In this model we have considered four different filter methods such as IG, GR, ReliefF and Correlation for feature selection in high dimensional data.

Information Gain (IG)

To indentify top ranked genes from pool of genes many researchers used information gain (IG) method. Help of this approach the difference between entropy and conditional entropy is of genes are gathered [32, 33].

\[
g(Y, X) = H(Y) - H\left(\frac{Y}{X}\right)
\]

H(Y) - Entropy of gene dataset Y (Ratio of uncertainty to predict the random variable)

H(Y/X)-Conditional Entropy (uncertainty with respect to the variable X)

\[
H(Y) = -\sum_{i} p_i \log_2 p_i
\]

\[
H(Y/X) = \sum_{x \in X} p(x) H(Y/X) = \sum_{x \in X} p(x)\frac{-\sum_{i} p_i \log_2 p_i}{p(x)}
\]

The rank order of individual gene is evaluated and from that high ranked feature genes are selected.

Gain Ratio (GR)

Information gain is suitable to identify the attributes having large no of values.C4.5 enhances the decision tree induction algorithm ID3. C4.5 is called gain ratio. Consider A is the set contains data sample with n different classes. So the information expected for classification is evaluated as

\[
R(A) = -\sum_{i} p_i \log_2 p_i
\]

\[
E(B) = \sum_{j} I(A) (D_{ij} + D_{ij} + ... + D_{nj}) / D \sum_{i=1}^{n}
\]

So Gain (A) =R (A)-E (B) (6)

Gain Ratio (B) =Gain (A)/Split info(A)

ReliefF

Instance based learning used by ReliefF to assign a weight to individual features. The weight of each helps to identify each from its different class values. According to the nearest hit /nearest miss (same instance class /opposite instance class) the weights of the features get updated. If it differentiates between inter classes then it receives high weight but in intra it maintains a same weight value. The value for nominal features the value defines either 1 or 0. 1 indicates values are different and 0 indicates values are same. So the nominal value of a feature varies between 0 & 1.

\[
Weight_{ReliefF} = \frac{Prob(\text{different value of features} / \text{different class})}{Prob(\text{different value of features} / \text{same class})}
\]

CFS

In 1999 Mark developed correlation based feature selection by focusing the relationship exists between individual attribute with respect to class. In CFS it indentifies high correlated dataset with its respected class by ignoring relation between them. By the feature selection techniques, irrelevant features are reduced and the high correlated genes can be indentified for classification.

C. Combinational Model

When the filter gene selection method is completed, high ranked feature subset are selected by IG, GR, ReliefF, Correlation feature selection techniques. The selected features
are taken as the top ranked ones as compared to other features. Not only considering all selected features for attribute evaluation and classification, may increase the computational cost not may it provide better classification accuracy. So to avoid redundant we again apply filter approach to enhance the classification accuracy.

D Attribute selection model and classification model
In the attribute selection model we have implemented SVM attribute selection evaluator. Addition to IG, GR, ReliefF, Correlation is combined with SVM were implemented to select the biomarker genes. In classification model the further classification of genes are done using LIBSVM classifier.

III. ALGORITHM OF THE PURPOSED MODEL
Input:
DS [FS1, FS2, FS3, FS4, ..., n]: n>=4
Ns <- Total no of feature in the subset.
Algorithm flow:
1. nFS<- ns // initialization of n feature
2. DS_c < - 10 CV (DS) // 10 cross validations
3. R IG = eval (IG, DS_cv) // Apply IG 10 cross validations
4. R GR = eval (GR, DS_cv) // Apply GR 10 cross validations
5. R RL = eval (RL, DS_cv) // Apply ReliefF 10 cross validations
6. R CFS = eval (CFS, DS_cv) // Apply CFS 10 cross validations
7. R mean = (R IG+ R GR+ R RL+ R CFS)/4 // Average feature positions
8. R mean = eval (R mean, DS_cv) // R mean ranking
9. Ranked Feature List (RFL)= [R IG, R GR, R RL, R CFS, R mean] // sorting Ascending order
10. For I = 1 to RFL
11. For J=1 to trunc( RFL_i/ns)
12. Do
13. S top ranked gene = extract (nF/ RFL_i)
14. S top ranked genes= 10 CV (S top ranked genes)
15. C_i (1, SVM) = eval (SVM, S_cv)
16. nFS <- NF+ns // feature update
17. nFS
18. end of for Loop
19. C (aux)= C (aux)+max(C)
20. End of second for
21. C (biomarker genes)=max (C (aux)) // biomarker gene with high statistical

This proposed model is supported by combination of 4 filter selection methods and machine learning classifiers (LIBSVM) for building a classification scheme that provide high performance. In this proposed model we have used IG, GR, ReliefF and Correlation for preliminary gene selection. For this we have used different attribute selection Evaluation and ranker selection tools available in Weka. The attribute evaluations evaluates clinical genes relevant to outcome based according to the filter methods. Then ranker ranks each genes on the basis of evaluation outcomes.

To indentify genes with high classification value we have considered filter and wrapper methods. The top ranked gene pool was created by considering top 5 percent ranked genes. With addition to filters ,SVM was employed as feature selection algorithm. At the end of the informative genes detected after SVM served as a data input for LIBSVM classifier to achieve classification accuracy. As the gene expression datasets are small in nature, for achieving high accuracy 10 cross validation was utilized. The above process was explained in the algorithm section.

IV. EXPERIMENT AND RESULT

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Proposed Method</th>
<th>IG SVM M</th>
<th>GRSVM M</th>
<th>ReliefFSVM VM</th>
<th>CORSVM M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leukemia</td>
<td>99.52</td>
<td>98.61</td>
<td>94.44</td>
<td>97.22</td>
<td>97.22</td>
</tr>
<tr>
<td>DLBC</td>
<td>100</td>
<td>100</td>
<td>97.87</td>
<td>95.74</td>
<td>100</td>
</tr>
<tr>
<td>Colon</td>
<td>91.26</td>
<td>90.32</td>
<td>85.48</td>
<td>87.1</td>
<td>87.1</td>
</tr>
<tr>
<td>Lung</td>
<td>100</td>
<td>100</td>
<td>98.96</td>
<td>98.96</td>
<td>98.96</td>
</tr>
<tr>
<td>Prostate</td>
<td>97.32</td>
<td>96.08</td>
<td>93.14</td>
<td>91.18</td>
<td>93.14</td>
</tr>
</tbody>
</table>

Table 2. Accuracy comparison

For the experimental analysis we have considered total 150 top most ranked genes from four filters by thinking that small no of genes are enough to produce high accuracy and the same is used as input for the SVM. After filter we understood that the top most genes are identified by different filter methods are almost similar, so top 5 genes are considered for classification purpose.

Table 2 represents the classification accuracy of different four hybrid methods individually (IGSVM, GRSVM, reliefFSVM, CorrelationSVM). From this table one can easily find out that comparison to all four filter methods IGSVM performed better. For some instance the IGSVM and Correlations achieved 100% for DLBCL dataset. Since various tools using default settings to achieve high accuracy by altering setting and selection is not possible. Mo et al. work out with prostate cancer with accuracy of 95.10% by using the randomized test.
Gene selection methods are used to determine biomarker genes from the microarray dataset. This study done with five various cancerous and noncancerous datasets are considered presented in the Table 1. The biomarker genes selected by our proposed model performed better as compared to individual, the detail of selective genes are mentioned in Table 3.

<table>
<thead>
<tr>
<th>Datasets Name</th>
<th>Selected Gene Description</th>
<th>Probe set Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Colon</td>
<td>F765 M76378_at</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F1423</td>
<td></td>
</tr>
<tr>
<td>Lung</td>
<td>F2968 M61906_at</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F4530 U45973_at</td>
<td></td>
</tr>
<tr>
<td>Prostate</td>
<td>F6185 37639_at</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F7067 40436_g_at</td>
<td></td>
</tr>
</tbody>
</table>

V. CONCLUSION

In this research study we have proposed a hybrid method with combination of four filter methods to select the biomarker gene by following a wrapper and classification method. After considering the best genes from all individual filters we can easily remove the irrelevant features from the microarray gene expression dataset which is high dimension in nature. A wrapper method is applied to eliminate the redundant natured genes out of 150 genes from the four filters. Finally after using LIBSVM classification tool we achieved 2 biomarker genes from three datasets such as Colon Cancer, Lung Cancer and Prostate Cancer. We demonstrated better performance with our proposed model with comparison with individual filter-wrapper classification approach. In this study we concluded that few biomarker genes are enough to achieve the classification accuracy from the microarray high dimension datasets. Even if we have detected some biomarker genes with high accuracy still we have to go though the functionality of the gene for accurate biomarker identification. This experiment is workload with small dimension datasets but it should need to be validating with large database.

VI. REFERENCE

Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE(pp. 5846-5849). IEEE.


