6.1 General

Tremendous development of DNA microarray technologies during past decade, plays very important role in the field of genomics and biomedical research. Microarrays have become a prospective tool for disease diagnosis or prognosis, by enabling the parallel observation of tens of thousands of gene expression at transcriptional level. Recognition of unique patterns of gene expression or identifying predictive genes correlated with a specific disease, is one of the most promising areas of genomics. Microarrays allow the detection of genetic changes attributable to a specific disease such as cancer by simultaneously analyzing the expression of thousands of genes in hopes of improving the specificity and sensitivity of these types of diagnostic assays. Subsequently, the increased application of microarray technology has provided a new dimension to gene expression based diagnostic pathology, specifically to cancer diagnostics.

One of the major objectives of the microarray data analysis, is to provide a generic approach for the cancer classification. However, cancer diagnosis and its classification based on gene expression is a great challenge due to the very high gene dimensionality and small sample size of microarray data. Also, the presence of redundant, irrelevant and noisy genes in the dataset degrades the computing efficiency as well as the classification accuracy of machine learning algorithms, particularly when samples are limited. Many machine learning and soft computing approaches have been implemented to get most relevant genes prior to cancer classification. The main objective of this research is to address the limitations in available cancer classification systems and contribute to the techniques of gene selection and cancer classification to improve on the selection of the optimal set of most relevant and non redundant genes for achieving better classification accuracy and making the analysis process fast and cost efficient. The main contributions of the thesis includes the development of

- An improved Binary Particle Swarm Optimization (iBPSO) based wrapper approach for
gene selection using DNA microarray gene expression data, is described in Chapter 3. The very characteristic of microarray data is high dimensionality (very large number of genes) and small number of tissue samples, which makes it easy to be trapped in a local optimum. The proposed iBPSO controls the problem of early convergence to the local optimum of traditional BPSO. The classification performance of each selected gene set is evaluated using three induction algorithms namely, Naive-Bayes (NB), k-Nearest Neighbor (k-NN) and Support Vector Machine (SVM) in this wrapper approach, by using five-fold cross-validation strategy.

- A case-based reasoning (CBR) framework with Tolerance Rough set-based gene selection for microarray cancer classification is presented in Chapter 4. The CBR approach is suitable for the microarray data because it requires much less domain knowledge than other types of learning approaches. This microarray based classifier is capable of self-learning and self-adapting to the unseen samples.

- A two phase hybrid model integrating Correlation-based Feature Selection (CFS) with improved-Binary Particle Swarm Optimization (iBPSO) for cancer classification, is reported in Chapter 5. This model selects a low dimensional set of prognostic genes to classify biological samples of binary and multi class cancers using Naive-Bayes classifier with 10-fold cross-validation. The proposed iBPSO also controls the problem of early convergence to the local optimum of traditional BPSO. The proposed model has been evaluated on 11 benchmark microarray datasets of different cancer types.

The work presented in this thesis for gene selection and cancer classification will be significantly useful in microarray data analysis. The methods presented in this thesis use supervised learning approaches which consider both challenges, high gene dimensionality and limited number of samples in microarray data analysis. The two proposed methods mainly focused on Binary Particle Swarm Optimization based wrapper gene selection approach, by improvising on inherent problems of local stagnation and inertia weight update. And the other one method utilized the CBR framework for cancer classification which is suitable for microarray data as limited number of samples present in it. The performance of methods are evaluated on binary
and multi-class cancer of different types, using stratified cross validation approach. The aim of this chapter is to highlight the main findings of the work carried out in this thesis and make suggestions for further research work in this thrust area of cancer classification based on microarray data analysis. The salient features of the developed cancer classification systems based on microarray data, have been shown in Table 6.1.

Table 6.1: Salient features of the methods proposed in the research work.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Technique</th>
<th>Salient features</th>
<th>Issues Addressed</th>
</tr>
</thead>
</table>
| 1.    | iBPSO (improved Binary Particle Swarm Optimization) | • Uses wrapper based gene selection approach.  
• iBPSO controls local stagnation problem of standard BPSO.  
• The classification performance of selected gene subset is evaluated using three learning approaches namely, NB, 1-NN and, SVM  
• iBPSO performed better in terms of classification accuracy and number of selected genes in comparison with standard BPSO. | • Problem of high gene dimensionality of microarray data.  
• Need for soft computing techniques for gene selection and cancer classification.  
• Better classification accuracy using small subset of predictive genes. |
| 2.    | TRFS-CBR (Tolerance Rough set-based Feature Selection with CBR Framework) | • Uses Case-based reasoning framework for cancer classification.  
• Suitable for microarray data, where limited domain knowledge is available.  
• Capable of gradually learning and improve its performance by already existing solved cases.  
• Experiments performed on seven cancer microarray datasets.  
• After retaining 20 cases in the initial case-base, the CBR framework has achieved upto 100% accuracy for Leukemia and mixed-lineage leukemia (MLL) datasets and 99.33% accuracy for SRBCT dataset.  
• System effectively reduced the gene dimensions upto < 3% of the total dimensions and achieved very high classification accuracy for all seven microarray datasets.  
• The comparative evaluation with some well-known feature extraction methods show that proposed CBR framework has attained best classification accuracy with higher Precision/Recall and Kappa ($\kappa$) values. | • Uses two-step gene selection procedure to extract most predictive genes related to a specific cancer type.  
• A microarray classifier capable of self-learning and self-adapting to the unseen samples.  
• Improve the prediction accuracy of CBR classifiers by reducing the number of attributes with minimal information loss. |
### Table 6.1: Salient features of the methods proposed in the research work. – continued from Page 124

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Technique</th>
<th>Salient features</th>
<th>Issues Addressed</th>
</tr>
</thead>
</table>
| 4.    | CFS-iBPSO\((\ddagger \omega)\)-NB (Correlation Feature Selection based improved Binary Particle Swarm Optimization) | • Uses hybrid framework of filter and wrapper feature selection approaches for gene selection.  
• Uses two phase hybrid model based on improved-Binary Particle Swarm Optimization (iBPSO) for Cancer diagnosis and classification.  
• Examined on 11 different types of Cancer microarray datasets and achieved 100% accuracy for seven datasets and with more than 92% for the remaining datasets.  
• Comparative performance evaluation of the proposed model is done with seven other benchmark methods and the model exhibits superior performance.  
• Selects small number (<1.5%) of highly relevant genes responsible for Cancer classification. | • Utilization of hybrid feature selection approach for gene selection in cancer microarray data.  
• An increasing inertia weight scheme \((\ddagger \omega)\) [27] is applied herein which controls the searching capability of the iBPSO algorithm in such a way that significantly improves the performance of the algorithm.  
• use of stratified 10-fold cross validation which is known for the best performance with lower bias and lower variance \([28, 29]\) in the estimation of test accuracy which provides more reliable and robust results. |

---

**6.2 Summary of Important Findings**

In Chapter 3, an improved Binary Particle Swarm Optimization (iBPSO) based wrapper approach is presented for gene selection in microarray gene expression data. Here three classifiers namely, Naive-Bayes, 1-NN and, SVM are used to evaluate the classification performance of the proposed method with 5-fold cross-validation. The proposed iBPSO also controls the problem of early convergence to the local optimum of traditional BPSO. For experimentation, six microarray cancer datasets are used to evaluate the performance of the proposed iBPSO method. These datasets represent binary and multi-class cancer classification problems. The experimental results show that the proposed method effectively improves the classification accuracy and selects a lower dimensional subset of genes. The iBPSO method achieved better optimization on the number of selected genes without affecting the classification accuracy by efficiently escaping from a local minima stagnation. Also, the results of Naive-Bayes classifier are better...
6.2: Summary of Important Findings

than the results of other two classifiers.

Limitations: Here, gene selection is performed using iBPSO wrapper approach which provides better and reliable prediction outcomes. However, in case of microarray datasets, using only a wrapper approach for gene selection has the limitations of slow computation speed and greater search complexity due to its very high dimensionality.

In Chapter 4, a novel CBR framework based on tolerance rough-set based feature selection (TRFS) is described for gene selection and cancer classification. The proposed framework relies on two step gene selection process to identify most relevant and non redundant genes which are attributable to the classification of a unknown cancer sample. FCBF (Fast Correlation-Based Filter) and TRFS select features based on single gene expression value which outputs a more interpretable case representation than other feature extraction methods such as PCA and LLE, which outputs features as a linear or non-linear combinations of all the input genes. The experimental results exhibit that the TRFS-CBR outperformed these feature extraction methods by a large margin in terms of accuracy and all other statistical measures. Also, the experimental results verify the capability of the proposed CBR system to self-learn, as new revised cases are included in the case base. So, the main advantages of our system are its robustness and efficiency to learn from previous stored cases. It achieves high classification accuracy using very small number of predictive genes. It also proved to be suitable for microarray data which have a small number of samples available.

Limitations: The limitation of our CBR system is the high classification cost in terms of response time. This is due to the fact that CBR is a lazy learning method which does all computations at time of classification of the unknown sample and as the case base extends with time, it requires more time to search the similar cases.

In Chapter 5, a two phase hybrid model for cancer classification is proposed, integrating Correlation-based Feature Selection (CFS) with improved-Binary Particle Swarm Optimization (iBPSO). In this proposed model, CFS reduces the dimensionality of the data by eliminating the irrelevant and redundant genes in the predictive gene pre-filtering phase. Then in gene optimization and cancer classification phase, iBPSO(↑w) makes use of this low dimensional gene subset to select an optimal subset of important genes with the help of Naive-Bayes classifier.
and 10-fold cross-validation which provides highest classification accuracy. We compared our model with seven popular methods from each category like Support Vector Machines, Random Forest, a filter (FCBF), a wrapper (standard BPSO and PSO-DT) and hybrid models (MBEGA and CFS-TCBPSO-1NN) on 11 benchmark cancer microarray datsets. Experimental results show that CFS-iBPSO(\( w \))-NB outperforms in terms of classification accuracy and number of selected genes in most cases. It attains 100% classification accuracy for seven datasets. CFS-iBPSO(\( w \))-NB effectively reduces the dimensionality by eliminating irrelevant and redundant genes and hence provides a low dimensional set of most important genes which are capable to achieve higher classification accuracy with much less complexity. Thus it could be an efficient tool for DNA microarray analysis.

**Limitations:** The proposed method CFS-iBPSO(\( w \))-NB is based on supervised learning approach which requires only labeled microarray data. However, in case of microarray data, availability of large amount of labeled tissue samples is very difficult.

### 6.3 Suggestions for Future Work

In this thesis, we have focused on the development of diagnostic models based on gene expression data which provide fast and reliable support in cancer diagnosis and its classification. These models can be applied to other microarray applications in clinical research, such as drug response analysis, patient’s survival and, toxicological studies. The main objective of this thesis is the selection of the optimal set of most relevant and non redundant genes for achieving better classification accuracy and making the analysis process fast and cost efficient. Further these results can be analyzed for the biological relevance so that this information would help biologists in providing biological interpretations of the outcomes found. The proposed models only utilized supervised learning approaches for cancer diagnosis and its classification and worked with labeled tissue samples only. However, most of the microarray datasets have limited number of labeled tissue samples. On the other hand, a large number of unlabeled microarray data that do not have proper clinical follow-up knowledge, are easier to obtain and can be used as additional information. Thus, the application of semi supervised learning approaches can also be investigated.