Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

International Conference on Mobility in Computing- ICMiC13, Organized by Mar Baselios College of Engineering and Technology during December 17-18, 2013 at Trivandrum, Kerala, India, pg.180 – 190

REVIEW ARTICLE

Review of Cancer Subtypes Prediction: Progress and Challenges

Gopika Rani G. U.[1], J. Jesu Vedha Nayahi[2]

[1,2]Computer Science and Engineering, Regional Centre of Anna University, Tirunelveli, Tamilnadu, India

[1] gugopikarani@gmail.com

Abstract— Microarray cancer data, organized as samples versus genes fashion, are being exploited for the classification of tissue samples into benign and malignant or their subtypes. They are also useful for identifying potential gene markers for each cancer subtype, which helps in successful diagnosis of particular cancer type. Nevertheless, small sample size remains a bottleneck to design suitable classifiers. Traditional supervised classifiers can only work with labelled data. On the other hand, a large number of microarray data that do not have adequate follow-up information are disregarded. This survey paper categorises, compares, and summarises the various approaches used in the field of cancer subtypes prediction in Gene Expression based Biological data mining. It defines the techniques used for cancer subtype prediction that includes the basic techniques like decision trees, rules, etc to the recent methodologies based on SVM. The association of feature selection methods with the classification methods is also included in this study which gives better performance in terms of accuracy and time complexity. Recent studies on cancer prediction introduce the scope of semi-supervised learning. This study presents methods and techniques in the field of semi-supervised cancer subtypes prediction incorporated with feature selection mechanisms and their issues. Compared to all related reviews, this survey proposes a novel approach to combine feature (gene) selection and transductive support vector machine (TSVM) where potential gene markers can be identified and TSVMs can improve prediction accuracy as compared to the standard inductive SVMs (ISVMs).

Keywords— Gene expression profiling; semi-supervised learning; gene selection; consistency; TSVM

I. INTRODUCTION

Biological data mining has become an essential part of a new research field called Bioinformatics. It emphasises on the genomic and proteomic data analysis. DNA sequences form the foundation of the genetic codes of all living organisms. All DNA sequences comprise basic building blocks called nucleotides. A gene usually comprises hundreds of nucleotides arranged in a particular order. Genome is the complete set of genes of an organism. Gene expression profiling is a technique used in molecular biology to query the expression of thousands of genes simultaneously. In case of tumor samples DNA labelled

with fluorophores (target) is prepared from a sample such as a tumor biopsy and is hybridized to the complementary DNA (cDNA) sequences on the gene chip. The chip is then scanned for the presence and strength of the fluorescent labels at each spot representing probe-target hybrids[25].

Cancer classification of different tumor types is of great importance in cancer diagnosis and drug discovery. A major challenge in clinical cancer research is the prediction of prognosis at the time of tumor discovery. Accurate prediction of different tumor types can help in providing better treatment and toxicity minimization on the patients. The advent of microarray technology has made it possible to study the expression profiles of a large number of genes across different experimental conditions. Microarray based gene expression profiling has shown great potential in the prediction of different cancer subtypes. The identification of tumor subtypes is based on established classification schemes such as the International Classification of Diseases published by the WHO which provides codes to classify diseases and a wide variety of signs, symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. For some types of cancer, these methods are unable to distinguish between subclasses. Leukemia consists of subtypes such as Acute lymphoblastic leukemia (ALL) and Acute myeloid leukemia (AML).Small round blood cell tumors (SRBCTM) include Ewings sarcoma (EWS) ,Nneuroblastoma (NB), Burkitt's lymphoma (BL) and Rrhabdomyosarcoma(RMS). Mixed-lineage leukemias (MLL) consists of ALL, MLL and AML. Diffuse large B-cell lymphomas (DLBCL) include subtypes such as Diffuse large B-cell lymphomas (DLBCL) and Follicular lymphoma(FL) .
The prediction of cancer subtypes includes the following issues:

A. Selection of relevant genes for classification

Before applying any mining technique, irrelevant attributes needs to be filtered. The main objectives of feature selection are to avoid overfitting and improve model performance and to provide faster and more cost-effective models. It is done using different feature selection techniques like wrapper, filter, embedded etc. In the filter approach the attribute selection method is independent of the data mining algorithm to be applied to the selected attributes and assess the relevance of features by looking only at the intrinsic properties of the data. In the wrapper approach the attribute selection method uses the result of the data mining algorithm to determine how good a given attribute subset is. In the embedded approach the attribute selection method is dependent of the data mining algorithm as well as look the intrinsic properties of the data. The selected features must be evaluated in terms of distance (euclidean distance measure), information (entropy, information gain, etc.), dependency (correlation coefficient), consistency (min-features bias) or classifier error rate.

B. Accurate Classification scenario for cancer subtypes

Various classifiers such as Rule based, Bayesian Networks, Decision tree, Nearest Neighbour, Artificial Neural Network, Rough set, Fuzzy logic based, Genetic Algorithm based, and Support Vector Machine(SVM) are used for cancer classification. The number of samples in microarray based cancer studies is usually small because microarray experiments are time consuming, expensive and limited by sample availability. Cancer classification using gene expression data usually relies on traditional supervised learning techniques, in which only labelled data (i.e., data from a sample with clinical follow-up) can be exploited for learning, while unlabeled data (i.e., data from a sample without clinical follow-up) are disregarded. Recent research in the area of cancer diagnosis suggests that unlabeled data, in addition to the small number of labelled data, can produce significant improvement in accuracy, a technique called semi supervised learning.

II. BACKGROUND

Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. In a few cases, such clinical heterogeneity has been explained by dividing morphologically similar tumors into subtypes with distinct pathogeneses. Key examples include the subdivision of acute leukemias, non-Hodgkin's lymphomas, and childhood "small round blue cell tumors" [tumors with variable response to chemotherapy that are now molecularly subclassified into neuroblastomas, rhabdomyosarcoma, Ewing's sarcoma, and other types]. For many more tumors, important subclasses are likely to exist but have yet to be defined by molecular markers.

A. Challenges in cancer subtype prediction

The cancer classification can be divided into two challenges: class discovery and class prediction. Class discovery refers to defining previously unrecognized tumor subtypes. Class prediction refers to the assignment of particular tumor samples to already-defined classes, which could reflect current states or future outcomes.

The data can be analyzed from many different viewpoints. The literature already abounds in studies of gene clusters discovered by unsupervised learning techniques[7]. Clustering is often done along the other dimension of the data. For example, each experiment may correspond to one patient carrying or not carrying a specific disease[9]. In this case, clustering usually groups patients with similar clinical records.

Classification is a technique used for discovering classes of unknown data. In the classification task, set of examples being mined is divided into two mutually exclusive and exhaustive sets called training set and test set. Basic classification process includes two phases such as training (where classification model is built from the training set) and testing(where the model is evaluated on the test set). Different approaches are used for classification that gradually made progress in terms of accuracy. These approaches are categorized as supervised, unsupervised and semi-supervised.

A known problem in classification specifically, and machine learning in general, is to find ways to reduce the dimensionality 'n' of the feature space F to overcome the risk of "over fitting". Data overfitting arises when the number 'n' of features is large (in our case thousands of genes) and the number of training patterns is comparatively small (in our case a few dozen patients). In such a situation, one can easily find a decision function that separates the training data (even a linear decision function) but will perform poorly on test data. Training techniques that use regularization avoid overfitting of the data to some extent without requiring space dimensionality reduction[22]. The identification of discriminant genes is of fundamental and practical interest. Research in Biology and Medicine may benefit from the examination of the top ranking genes to confirm recent discoveries in cancer research or suggest new avenues to be explored.

Performing feature selection in large dimensional input spaces therefore involves greedy algorithms. Among various possible methods, feature-ranking techniques are particularly attractive. A fixed number of top ranked features may be selected for further analysis or to design a classifier. Alternatively, a threshold can be set on the ranking criterion. Only the features whose criterion exceeds the threshold are retained. In the spirit of Structural Risk Minimization it is possible to use the ranking to define nested subsets of features $F_1 \subset F_2 \subset \cdots \subset F$, and select an optimum subset of features with a model selection criterion by varying a single parameter: the number of features.

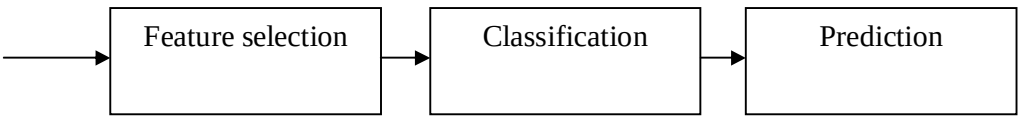

Fig 1 Overall architecture of cancer subtype prediction

B. Organization of the paper

The rest of the paper is organized as follows: In section III points out the methods of cancer prediction and discuss the advantages and disadvantages of each technique. Section A describes the previous works that combines cancer subtype prediction with gene selection, Section B discusses the cancer subtype prediction without gene selection, Section C gives the trends in semi-supervised learning approaches and Section D describes various approaches on feature selection that used for gene selection. Its present suggestions, new proposals and future directions in Section IV that is inferred from the survey sections. The study can be concluded in Section V.

III. METHODS FOR CANCER SUBTYPE PREDICTION

The advent of microarray technology has made it possible to study the expression profiles of a large number of genes across different experimental conditions. Microarray-based gene expression profiling has shown great potential in the prediction of different cancer subtypes. Small sample size remains a bottleneck in obtaining robust and accurate prediction models [13], [15]. The number of samples in microarray based cancer studies is usually small because microarray experiments are time consuming, expensive, and limited by sample availability.

Cancer classification using gene expression data usually relies on traditional supervised learning techniques, in which only labelled data can be exploited for learning, while unlabeled data are disregarded. New methods in which unlabeled data, in addition to the small number of labelled data, can produce significant improvement in accuracy, a technique called semisupervised learning.

Cancer classification using microarray data poses another major challenge because of the huge number of features (genes) compared to the number of examples (tissue samples). This is an important problem in machine learning which is known as feature selection [6]. Successful gene identification involves 1) dimension reduction to reduce computational cost; 2) reduction of noise to increase classification performance; and 3) identification of more interpretable features. Only a small number of genes in the microarray data consisting of thousands of genes show strong correlation with the target phenotypes. Only a few small selected genes have their biological relationship with the target diseases.

The following sections describes various approaches used for the cancer subtype prediction as well as research on techniques such as semi-supervised learning and feature selection that can be integrated with the previous methods to increase accuracy and avoid overfitting problem.

A. Cancer Subtype Prediction with Gene Selection

The combination of gene selection and cancer classification can give a better accuracy when the relevant genes are only used for classification. The various methods for feature selection and classifications are used with associated pros and cons. The common advantage that we can infer from these approaches includes the gradual increase in accuracy.

In 2002, Guyon published a method for Gene selection for cancer classification using Support Vector Machine[1]. It includes a Recursive Feature Elimination(RFE) to select subset of features without redundancy. It is combined with SVM to classify informative patterns to cancer classes. Even though it increase the accuracy, it could not be accurate when there are large number of features over number of training samples. This drawback is based on the RFE that will eliminate one feature at a time based on the ranking performed on the whole feature set.

Another variation in this category comes in 2007 which replaced other data mining techniques for feature selection and classification. Genetic Algorithm(GA) based method is used for the feature selection in which t-statistics(t-GA) is calculated for selecting relevant genes[2]. The classification here used is decision tree based and offer more accuracy. At the same time stability of the classification is obtained even the number of genes is changed. It is also reproducable in gene selection because it gives consistent set of genes whenever the parameters applied on the algorithm are changed. But decision tree based approaches limit the dimensionality of the problem. Based on the values of 'p' (no. of features to be selected with upper t-GA) , the decision tree varies.

In a research article in BMC Bioinformatics,2011 the possibility to use single genes to predict cancer[3]. In this the most powerful genes with univarient class discrimination ability is identified. It use simple classification rules for class prediction using those single genes. Even it can reduce the overfitting of data, it is not applicable for problems without highly differentially expressed genes. So in this case, need to include multiple genes.

Mining and integrating reliable decisions rules are more powerful gene selection algorithm for imbalanced data sets [4]. Here, Skewed gene selection algorithm that introduces a weighted metric for the gene selection. The extracted genes are paired as decision rules to distinguish both classes, with these decision rules. It is integrated into an ensemble learning framework by majority voting to recognize test examples. It will avoid data normalization and classifier construction and give a specific design for imbalanced gene expression datasets to acquire better recognition capabilities. The limitations should be noted that mining of decision rules is quite time consuming and it only deals with binary class skewed gene expression rather than multiclass.

A related work in 2011 provides application of Artificial Neural Network in DNA classification which uses feature extraction and redundancy analysis[5]. The extracted genes undergoes ANN based modelling. The SVM classifier is used for getting result. It will increase the accuracy but the application range of the classification method is not wide enough. It is necessary to optimize the feature extraction and classification method.

In 2012, a data integration model for cancer subtype prediction using Kernel Dimensionality Reduction-SVM is suggested[6]. It uses a kernel based classification methods which is an extension of SVM with KDR. It also employs an integration process of patients data types such as clinical data and micro array DNA. This accuracy enhanced method has the limitations that includes late integration in case of unavailability of patient's data, and problems on integration of different data set which is difficult because of dimensional problems.

B. Cancer Subtype Prediction Without Gene Selection

There are various classification strategies that improve accuracy but not accompanied with the gene selection procedure. It includes different classification as well as clustering techniques that is used for cancer subtype prediction. In 2000,a technique used to identify distinct types of large B-cell lymphoma is proposed that incorporate gene expression profiling using hierarchical clustering based on gene expression signature[7]. Here the genes to represent the subtypes are selected. These genes are again hierarchically clustered to obtain the result. It takes advantages in prognosis and to find different features of patients that influence their survival. The main drawback is that it takes patient's age, performance status, extent and location of disease. It can only useful for patients with low clinical risk.

A gene signature based method, for identifying subtypes and subtype-specific drivers in cancer with an application to medulloblastoma is used to identify the cancer driver[8]. It is an iterative approach that alternates between kernel based gene expression clustering and gene signature selection. Here, top ranked driver candidates are found to be enriched with

known pathways in certain type. It use CNA (copy number aberrations) events that initiate and drive the cancer. Advantages includes the identification of specific CNA driver and it allows functional interpretation of the cancerous process or pathways. The disadvantages includes the existence of outliers, issues in selecting the correct clustering algorithm, challenges in determining number of subtypes etc. It does not give much emphasis on subtype specificity.

A report in Science Magazine in 1991 introduces the issues such as class discovery and class prediction by gene expression monitoring[9] . It gives a new strategy for cancer prediction without medicinal knowledge and with improved prediction. It also forward the issues such as the accurate prediction that depends on the strategy used. After that many selection strategies for improving accuracy is proposed by different authors. Among them Classification and diagnostic prediction of cancer using ANN and gene expression profiling[10] need more attention , which uses Gene signature based ANN classifier and use blinded samples that are previously not used in training set. It includes ranked order gene identification with high sensitivity and specificity. But it can't achieve in case of large arrays of sample data. Mining top-k covering rule group gives faster classification using association mining algorithm[11]. Here the rule group are generated and the classification performed according to the rules. It eliminates excessive number of redundant rules. But it is less specific.

An improved version of ensemble machine learning algorithm that improve accuracy by means of bagging, boosting, and arcing methods can be used as sequential self optimization structure[12]. But there is no selection of influential attributes or genes and no attribute precedence given for base model. For mining cancer data a classification based on Discrete Particle Swarm optimization and rule pruning can be used in which decision rules are discovered through training data[13].It can achieve more performance than PSO. Rule pruning address the issue of overfitting the training data by removing the irrelevant terms. But rule pruning is effective in predicting only common types of cancers. DPSO depends on constants randomly used for position adjustments.

The improvement on multiobjective clustering through SVM with application to gene expression data is used to monitor expression profile of a large number of genes across different experimental conditions simultaneously in improved manner[14]. Here, the combination of multiobjective fuzzy clustering scheme with SVM that will produce set of non-dominated solutions and give high confidence points using fuzzy voting technique. This effective method for gene expression data is sensitive to algorithmic parameters. Based on the breast cancer a multi-attributed Lens Recursive partitioning algorithm is used to overcome the misclassification error by a recursive partitioning algorithm where a single attribute is selected from candidate attributes to split a dataset using information measures[15]. Multi-attributed lens which weighs all numeric attributes simultaneously. A lens is generated using a core vector from a farthest pair of the same class instances. Data is partitioned into two regions- the outside and the inside lens. All instances in the outside lens are marked as opposite classes to the core vector. It is better than c4.5 and it consider multiple attribute at a time. But it need more effective detection classifier where attribute must eliminate using negative lens and group using positive lens.

C. Semi Supervised Learning Approaches

In case of cancer classification, small sample size remains a bottleneck in obtaining robust and accurate prediction models. The number of samples in microarray based-cancer studies is usually small because microarray experiments are time consuming, expensive, and limited by sample availability.

Recent research in the area of cancer diagnosis suggests that unlabeled data, in addition to the small number of labelled data, can produce significant improvement in

accuracy, a technique called semi supervised learning [24]. Indeed, semisupervised learning has proved to be effective in solving different biological problems including protein classification, prediction of transcription factor–gene interaction, and gene- expression based cancer subtype discovery. Major research on extending support vector machines (SVMs) to handle semi labelled data is based on the following idea: solve the standard inductive SVM (ISVM) while treating the unknown labels as additional optimization variables. By maximizing the margin in the presence of unlabeled samples, one can learn the decision boundary that traverses through low density regions while respecting labels in the input space. In other words, this approach implements the cluster assumption for semi supervised learning, that samples in a data cluster have identical labels. The idea was first introduced under the name of transductive SVM, but since it learns an inductive rule defined over the entire input space, the approach is referred to as semisupervised SVM (S^3VM). Each cluster of samples is assumed to belong to one data class. Thus, a decision boundary is defined between clusters. A variety of semisupervised techniques have been proposed and many successful algorithms directly or indirectly assume high density within class and low density between classes, and can fail when the classes are strongly overlapping[24], [23]. This can be illustrated by comparing the well-known SVMs to their semisupervised extension, transductive SVM, progressive TSVM algorithm (PTSVM) [20], transductive SVMs (TSVMs), and semisupervised SVMs (S^3VMs). TSVMs and S3VMs are iterative algorithms that use SVMs to gradually search a reliable hyperplane exploiting both labeled and unlabeled samples in the training phase. The various approaches in semi supervised learning can be effectively discussed for further improvement in the field.

In 2002,semi supervised learning with progressive transductive support vector machine is used to extend TSVM to handle different class distribution[20].It employs the pair wise labelling and dynamic adjustment using unlabeled data. It can give better testing result as well as the generalization performance. The limitation of this method is that the progressive movement of hyperplane in one direction cause misclassification. It is not suited for large data. Semi- supervised classification by Low density Separation method is to exploit the cluster assumption for successful semi-supervised learning[21]. It gives three algorithms: deriving graph based distances that emphasize low density regions between clusters followed by training a standard SVM, optimizing the Transductive SVM objective function, this places the decision boundary in low density regions, by gradient descent and combining first two to make maximum use of the cluster assumption. It will increase the accuracy, but fails when there is overlapping between the classes. It doesn't provide gradual improvement on labelling the unlabeled data. Manifold Regularization includes a geometric framework for learning from labeled and unlabeled examples[22]. It is aimed to exploit the geometry of the probability distribution that generates the data and incorporates it as an additional regularization term. In manifold learning the attempt is to use the geometry of the probability distribution by assuming that its support has the geometric structure of a Riemannian manifold. It will reproduce Kernel Hilbert Spaces(RKHS). This lead to the class of kernel based algorithms for classification and regression. It can have better usage of unlabeled data. But it is efficient for low dimensional data only and in case of non-linear problem, it give an approximation only.

The technique using Laplacian Normalization for graph semi-supervised learning investigates the effect of laplacian regularization in multi-class transductive learning on graphs[23]. Here the generalization bounds are derived using geometric properties of the graph. It uses the graph laplacian matrix normalization. It will overcome the limitations of the standard degree-based normalization method and doesn't cause inferior generalization performance. Graph regularization and normalization increase the complexity of learning process. Various optimization techniques for semi-supervised support vector machine are identified[24]. It

introduced the technique to optimize the semi-supervised learning using semi-supervised support vector machines(S^3VMs) based on applying the margin maximization principle to both labeled and unlabeled examples. Their formulation leads to a non-convex optimization problem. It will yield good performance on textual data sets than manual labelling and improve generalization. But it is not suited for structured output problems.

D. Gene(Feature) Selection Techniques

There are various approaches from 1997 when the first paper on feature selection is published[16]. Feature selection can be stated as follows: P is the total number of instances, N denotes the total number of features, M stands for the number of relevant/selected features, S denotes a subset of features, $f_1, \ldots, f_M$ are the M features, m denotes the number of different class labels, and C stands for the class variable.

It addresses two issues –problem of selecting relevant features and relevant examples. It forwarded the filter, wrapper and embedded approaches on feature selection. It is applicable for selecting relevant features by eliminating redundancy. But there is a need to evaluate the relevance of feature selection based on various measures. The selected features must be evaluated based on some measure. The evaluation functions can be divided into five categories: distance, information (or uncertainty), dependence, consistency, and classifier error rate. In the following, briefly discuss each of them.

1) Distance Measures: It is also known as separability, divergence, or discrimination measure. For a two class problem, a feature f_i is preferred to another feature f_j if f_i induces a greater difference between the two-class conditional probabilities than f_j ; if the difference is zero then f_i and f_j are indistinguishable.

2) Information Measures: These measures typically determine the information gain from a feature. The information gain from a feature f_i is defined as the difference between the prior uncertainty and expected posterior uncertainty using f_i . Feature f_i is preferred to feature f_j if the information gain from feature f_i is greater than that from f_j.

3) Dependence Measures: Dependence measures or correlation measures quantify the ability to predict the value of one variable from the value of another variable. Correlation coefficient is a classical dependence measure and can be used to find the correlation between a feature and a class variable. If the correlation of feature f_i with class variable *C* is higher than the correlation of feature f_j with *C*, then feature f_i is preferred to f_j . A slight variation of this is to determine the dependence of a feature on other features; this value indicates the degree of redundancy of the feature. All evaluation functions based on dependence measures can be classified as distance and information measures. But, these are still kept as a separate category because, conceptually, they represent a different .

4) Consistency Measures: This type of evaluation measures are characteristically different from other measures because of their heavy reliance on the training dataset and use of Min-Features bias in selecting a subset of features . Min-Features bias prefers consistent hypotheses definable over as few features as possible. These measures find out the minimal size subset that satisfies the acceptable inconsistency rate, that is usually set by the user.

The above types of evaluation measures are known as "filter" methods because of their independence from any particular classifier that may use the selected features output by the feature selection method.

5) Classifier Error Rate Measures: In contrast to the above filter methods, classifier error rate measures are called "wrapper methods", i.e., a classifier is used for evaluating feature subsets. As the features are selected using the classifier that later uses these selected features in predicting the class labels of unseen instances, the accuracy level is very high although computational cost is rather high compared to other measures.

The relevant genes can be selected that eliminate redundant features, but relevance of the features must be evaluated on various measures[16]. It need to compile the algorithm each time for classification.

Feature selection can effectively use consistency as the evaluation measure[17] . It calculate the inconsistent rate that can be used to find the optimal subset of features that is consistent. A feature subset is said to be inconsistent if there exist at least two instances with same feature value but with different class labels. It will increase the overall accuracy of classification while the data size is reduced. The comprehensibility is also improved. But it does not incorporate any search bias with regards to a particular classifier. It forwards five

search strategies-FocusM, ABB, SetCover, LVF(Las Vegas Filter), QBB(Quick Branch and Bound).If time is not an issue, then FocusM and ABB are preferable because they ensure smallest consistent subset. In addition, if the user has knowledge of *M* then if *M* is small FocusM is preferable otherwise ABB is chosen. In the absence of any a priori knowledge about *M* both can be run simultaneously until any one terminates. In the usual case of limited computing time a user is best off choosing from LVF, SetCover and QBB. LVF is suitable for quickly producing small (yet not small enough) consistent subsets. SetCover is suitable if it is known that features are not correlated.

The problem is to select those input features that are most predictive of a given outcome. As a solution we can use a crisp and fuzzy set-based methodology named Rough Set Attribute reduction(RSAR) that provides a filter based tool by which knowledge may be extracted from a domain in a concise way[18]. RSAR requires no additional parameters to operate other than the supplied data. It can use the technique called feature grouping. The limitations include the cost of entropy measure which is costlier than the dependency evaluation. The most basic solution to locating a subset is to simply generate all possible subsets and retrieve those with a maximum entropy degree. Obviously, this is an expensive solution to the problem and is only practical for very simple data sets. Most of the time only one reduct is required, so all the calculations involved in discovering the rest are pointless. It is an important factor when large data sets are processed.

In 2012, a new feature selection algorithm is proposed to overcome the drawback of ranking gene selection whereby a weakly ranked gene that could perform well in terms of classification accuracy with an appropriate subset of genes[19]. In this technique first the genes are divided into subsets of relatively small size (say h). Then the informative smaller subsets of genes of size r<h) from a subset are selected and merged with the chosen genes with another gene subset(of size r) to update the gene subset. The process is repeated until all the subsets are merged into one informative subset. It can achieve high classification accuracy. Also the observed genes have biological relevance. At the same time the classification accuracy depends on 'r'-size of the subset. Another limitation is the increase in processing time for block reduction.

IV. TRANSDUCTIVE SVM AND CONSISTENCY BASED FEATURE SELECTION

Major research on extending support vector machines (SVMs) to handle semilabelled data is based on the following idea: solve the standard inductive SVM (ISVM) while treating the unknown labels as additional optimization variables. A variety of semisupervised techniques have been proposed and many successful algorithms directly or indirectly assume high density within class and low density between classes, and can fail when the classes are strongly overlapping.

To overcome the limitation of small sample size TSVMs are used which are basically iterative algorithms that gradually search the optimal separating hyperplane in the feature space with a transductive process that incorporates unlabeled samples in the training phase. Unlike the selection procedure of transductive samples in traditional TSVMs, the selection of transductive samples is done through a process of filtering the unlabeled samples. Unlabeled samples that fall into the margin will have richer information to find a better separating hyperplane. These samples are selected and working set is updated and retrained the TSVM for each iteration. This procedure improves the generalization capability of the classifier. Gradually, the separating hyperplane will move to a finer position in subsequent iterations.

Cancer classification using microarray data poses another major challenge because of the huge number of features (genes) compared to the number of examples (tissue samples). Successful gene identification involves dimension reduction to reduce computational cost, reduction of noise to increase classification performance and identification of more

interpretable features. Only a small number of genes in the microarray data consisting of thousands of genes show strong correlation with the target phenotypes. Only a few small selected genes have their biological relationship with the target diseases.

A new feature selection method incorporated with a consistency based forward greedy gene selection procedure can be used to find gene marker for each of the subtypes of the cancer. It is aimed to develop a classification system by identifying potential gene markers and subsequently applying the proposed TSVM technique on the selected genes for the classification of human cancer. A forward greedy reduction algorithm was exploited to identify the gene markers. It is to obtain an effective technique compared with the LDS and ISVM on the basis of overall average accuracy

V. CONCLUSIONS & FUTURE DIRECTIONS

The present study was designed to address the challenges in the field of cancer subtype prediction and the different measures of gene selection used. Small sample size cause a problem in gene-expression based- outcome prediction for human cancers. In TSVM algorithms, the transductive samples are selected on the basis of geometric analysis of the feature space, and only support vector-like samples that contain the richest information are included in the training set. In particular, the proposed TSVM is an iterative procedure that defines the hyperplane according to a transductive process that integrates unlabeled samples together with the training samples. It also gives a consistent set of genes with higher discriminatory power using consistency based forward greedy algorithm. Moreover, as a scope of future work, we can plan to apply fuzzy rough set theory to find more relevant gene markers and introduce fuzzy set theory in transductive/semisupervised learning to improve the performance of cancer subtype prediction. It can also accompany with other feature selection strategy that give relevant set of features in lesser processing time.

REFERENCES

[1] Isabelle Guyon, Jason Weston, Stephn Barnhill, *Gene Selection for Cancer Classification using Support Vector Machines*, Journal-Machine Learning, 46, 389–422, 2002

[2] Jinn-Yi Yeh, Tai-Shi Wu, Min-Che Wu, Der-Ming Chang, *Applying Data Mining Techniques for Cancer Classification from Gene Expression Data*, IEEE International Conference on Convergence Information Technology,2007

[3] Xiaosheng Wang and Richard Simon, *Microarray-based cancer prediction using single Genes*, ,Research Article-BMC Bioinformatics,2011

[4] Hualong Yu, Jun Ni, Yuanyuan Dan, Sen Xu, *Mining and Integrating Reliable Decision Rules for Imbalanced Cancer Gene Expression Data Sets*, Journal- Tsinghua Science And Technology Issnll1007-0214 Dec 2012

[5] Qingda Zhou Qingshan Jiang Shenzhen ,Dan Wei, *A New Method for Classification in DNA Sequence*, The 6th International Conference on Computer Science & Education, IEEE 2011

[6] Ito Wasito Aulia N. IstiqlalIndra Budi, *Data Integration Model for Cancer Subtype Identification using Kernel Dimensionality Reduction-Support Vector Machine (KDR-SVM)*, IEEE conference 2012

[7] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Article-Nature Magazine, Feb 2000

[8] Peikai Chen and Y. S. Hung, Tsz-Kwong Man and Ching, Yubo Fan and Stephen T.-C. Wong, *A gene signature based method for identifying subtypes and subtype-specific drivers in cancer with an application to medulloblastoma* , IEEE Conference 2012

[9] T. R. Golub, D. K. SlonimP., Tamayo, C. Huard, *Molecular Classifcation of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, Report-Science Magazine,1991

[10] Javed Khan, Jun S. Wei, Markus Ringnér, *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*, Article-Nature Medicine, June 2001

[11] Gao Cong , Kian-LeeTan, Anthony K.H.Tung , Xin Xu, *Mining Top-k Covering Rule Groups for Gene Expression Data*, ACM,2005

[12] Ching Wei Wang , *New Ensemble Machine Learning Method for Classification and Prediction on Gene Expression Data*, 28th IEEE EMBS Annual International Conference,IEEE,2006

[13] Yao Liu and Yuk Ying Chung, *Mining Cancer data with Discrete Particle Swarm Optimization and Rule Pruning*, ,IEEE Conference 2011

[14] Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay, *Improving Multi-objective Clustering through Support Vector Machine: Application to Gene Expression Data*, Springer,2011

[15] Charoenchai Sirisomboonrat, Krung Sinapiromsaran, *Breast Cancer Diagnosis using Multi-Attributed Lens Recursive Partitioning Algorithm*, Tenth International Conference on ICT and Knowledge Engineering,2012

[16] Avrim L. Bluma, Pat Langley, *Selection of relevant features and examples in machine learning*, Elsevier,1997

[17] Manoranjan Dash , Huan Liu , *Consistency-based search in feature selection*, Elsevier,2003

[18] Richard Jensen And Qiang Shen, *Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough-Based Approaches*, IEEE Transactions On Knowledge And Data Engineering, December 2004

[19] Alok Sharma, Seiya Imoto, And Satoru Miyano, *A Top-r Feature Selection Algorithm for Microarray Gene Expression Data*, IEEE/ACM Transactions On Computational Biology And Bioinformatics, May/June 2012

[20] Yisong Chen , Guoping Wang, Shihai Dong, *Learning with progressive transductive support vector machine*, Elsevier,2002

[21] Olivier Chapelle, Alexander Zien, *Semi-Supervised Classification by Low Density Separation*, 10th international Workshop on Artificial Intelligence ,2005

[22] Mikhail Belkin, Partha Niyogi,Vikas Sindhwani, *Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples*, Journal of Machine Learning Research 2006

[23] Rie Johnson, Tong Zhang, *On the Effectiveness of Laplacian Normalization for Graph Semi-supervised Learning*, Journal of Machine Learning Research ,2007

[24] Olivier Chapelle, Vikas Sindhwani, Sathiya S.Keerthi, *Optimization Techniques for Semi-Supervised Support Vector Machines*, Journal of Machine Learning Research , 2008

[25] Mingguang Shi and Bing Zhang, *Semi-supervised learning improves gene expression-based prediction of cancer recurrence*, Publication of Wikipedia,2011

