Feature Selection of High Dimensional Big Data of Gene Expression for Cancer Dataset

Prem Kumar Chandrakar[1], A. K. Shrivas[2]
[1] Dept. of Computer Science, Mahant Laxminarayan Das College, Raipur (C.G.) India.
[2] Dept. of CS & IT, Dr. C.V. Raman University, Bilaspur (C.G.) India.

ABSTRACT
Feature selection is an essential data preprocessing technique for such high-dimensional data classification tasks. Traditional dimensionality reduction approach falls into two categories: Feature Extraction (FE) and Feature Selection (FS). The microarray technology has capability to determine the levels of thousands of gene simultaneously in a single experiment. The major challenge to analyze gene expression data, with a large number of genes and small samples, is to extract disease-related information from a massive amount of redundant data and noise. Analysis of gene expression is important in many fields of biological research in order to retrieve the required information. As time progresses, the illness in general and cancer in particular have become more and more complex and complicated, in detecting, analyzing and curing. We know cancer is deadly disease. Cancer research is one of the major area of research in medical field. Predicting precisely of different tumor types is a great challenge and providing accurate prediction will have great value in providing better treatment to the patients. To achieve this, data mining algorithms are important tools and the most extensively used approach to achieve important feature of gene expression data and plays an important role for gene classification. Gene expression profiles, which represent the state of a cell at a molecular level, has greatpotential as a medical diagnosis tool. But compared to the number of genes involved, available trainingdata sets generally have a fairly small sample size for classification. These training data limitationsconstitute a challenge to certain classification methodologies. Feature selections techniques can be usedto extract the marker genes which influence the classification accuracy effectively by eliminating the unwanted noisy and redundant genes. One of major challenges is to discover how to extract useful information from huge datasets. Gene selection, eliminating redundant and irrelevant genes, has been a key step to address this problem.This paper presents a various of feature selection techniques that have been employed in micro array data based cancer classification and presents recent advances in the machine learning based gene expression data analysis with different feature selection algorithms.

Keyword–Gene Expression, Cancer Classification, Feature selection

I INTRODUCTION

Feature selection is an active research area in pattern recognition, statistics, and data miningcommunities. The main idea of feature selection is to choose a subset of input variables byeliminating features with little or no predictive information. Feature selection can significantlyimprove the comprehensibility of the resulting classifier models and often build a model thatgeneralizes better to unseen points. Further, it is often the case that finding the correct subset ofpredictive features is an important problem in its own right. For example, physician may make adecision based on the selected features whether an expensive surgery is necessary for treatmentor not.

II DNA MICROARRAY

Microarray technology is a developing technology used to study the expression of many genes atonce. It involves placing thousands of gene sequences in known locations on a glass slide calleda gene chip. A sample containing DNA or RNA is placed in contact with the gene chip.Complementary base pairing between the sample and the gene sequences on the chip produceslight that is measured. Areas on the chip producing light identify genes that are expressed in the sample.

Microarray technology provided an opportunity for the researchers to analyze thousands of geneexpression profiles simultaneously that are relevant to different fields including medicineespecially cancer. The categorization of patient gene expression profile has become a commonstudy in biomedical research. The real problem is managing microarray data with its dimension.Since the dimension of microarray is large, classifying and handling the algorithms becomes toocomplex to study the gene expression characteristics. Due to the presence of more improperattributes in the dataset, the accuracy of the classification algorithm also gets affectedsignificantly. The aim of feature selection algorithm is to isolate the most important featuresfrom the microarray data to minimize the feature space in order to improve the accuracy of theclassification.

A microarray gene expression data set can be represented in a tabular form, in which each rowrepresents to one particular gene, each column to a sample or time point, and each entry of thematrix is the measured expression level of a particular gene in a sample or time point,respectively.

DNA microarrays are created by robotic machines that arrange large amounts of hundreds orthousands of gene sequences on a single microscope slide. Researchers have a database of over 40,000 gene sequences that they can use for this purpose. When a gene is activated, cellularmachinery begins to copy certain segments of that gene. The resulting product is known asmessenger RNA (mRNA), which is the body's template for creating proteins. The mRNAproduced by the cell is complementary, and therefore will bind to the original portion of theDNA strand from which it was copied.

To determine which genes are turned on and which are turned off in a given cell, a researchermust first collect the messenger RNA molecules present in that cell. The researcher then labelseach mRNA molecule by using a reverse transcriptase enzyme (RT) that generates acomplementary cDNA to the mRNA. During that process fluorescent nucleotides are attached tothe cDNA. The tumor and the normal samples are labeled with different fluorescent dyes. Next, the researcher places the labeled cDNAs onto a DNA microarray slide. The labeledcDNAs that represent mRNAs in the cell will then hybridize – or bind – to their syntheticcomplementary DNAs attached on the microarray slide, leaving its fluorescent tag. A researchermust then use a special scanner to measure the fluorescent intensity for each spot/areas on themicroarray slide.

If a particular gene is very active, it produces many molecules of messenger RNA, thus, morelabeled cDNAs, which hybridize to the DNA on the microarray slide and generate a very brightfluorescent area. Genes that are somewhat less active produce fewer mRNAs, thus, less labeledcDNAs, which results in dimmer fluorescent spots. If there is no fluorescence, none of themessenger molecules have hybridized to the DNA, indicating that the gene is inactive.Researchers frequently use this technique to examine the activity of various genes at differenttimes. When co-hybridizing Tumor samples (Red Dye) and Normal sample (Green dye)together, they will compete for the synthetic complementary DNAs on the microarray slide. As aresult, if the spot is red, this means that that specific gene is more expressed in tumor than innormal (up-regulated in cancer). If a spot is Green that means that that gene is more expressed inthe Normal tissue (Down regulated in cancer). If a spot is yellow that means that that specificgene is equally expressed in normal and tumor.

III RELATED WORK

Cancer is one of most deadly dieses and lot of people die world- wide because of cancer. As per the WHO statistics in 2018 more than 20 million new cases were identified and around 9.6million cancer related death occur. Globally, about 1 in 6 deaths is due to cancer.The number of new a case is expected to rise by about 70%over the next2 decades (source: WHO 2018). It has been identified long ago that cancer occurs because of genedisorder. Gene expression is nothing but level of production of protein molecules defined by agene. Monitoring of gene expression is one of most fundamental approach in genetics. Thetechnique for measuring gene expression is to measure the mRNA instead of protein, becausemRNA sequences hybridize with their complementary RNA or DNA sequence while thisproperty lacks in protein. The DNA arrays are novel technologies that are designed to measuregene expression of tens of thousands of genes in a single experiment. Gene expression datausually contain a large number of genes (in thousands) and a small number of experiments (indozens). In machine learning terminology, these data sets are usually of very high dimensionswith undersized samples. The purpose of Gene selection is to find a set of genes that bestdiscriminate biological sample of different types. The selected genes are “biomarkers,” and theyform a “marker panel” for analysis. For analyzing the marker panel rank based scheme suchinformation gain was used. It was observed that the information gain with large group was notaccurate, therefore in paper**(Zhu, Wang, Yu, Li, & Gong, 2010)**they proposed model-based approach to estimate the entropy onthe model, instead of on the data themselves. Here, they used multivariate Gaussian generativemodels, which model the data with multivariate normal distributions.

IV FEATURE SELECTION METHOD

There are two types of feature selection methods have been studied: filter methods **(Langley, Flamingo, & Edu, 1994)** andwrapper methods**(Kohavi & John, 1997)**.Filter methods are essentially data preprocessing or data filtering methods. Features are selectedbased on the intrinsic characteristics that determine their relevance or discriminative powers withregard to the target classes.In wrapper methods, feature selection is “wrapped” around a learning method: the usefulness ofa feature is directly judged by the estimated accuracy of the learning method. Wrapper methods typically require extensive computation to search for the best features.

(a) Basic feature selection algorithm

(i) Input:

S - Data sample f with features X, |X| = n

J - Evaluation measure to be maximized

GS – successor generation operator

(ii) Output:

Solution – (weighted) feature subset
L: = Start Point(X);
Solution: = {best of L according to J };

(iii) Repeat

L: = Search Strategy (L, GS (J), X);
X':= {best of L according to J};
If J (X') =J (Solution) or (J (X') =J (Solution) and |X'| < |Solution|) then
Solution: =X';

(iv) Until Stop (J, L).

The discriminating criteria are being used by filter method for feature selection. The correlationcoefficient or statistical test like t-test or f-test is used to filter the features in the filter featureselection method.Many interesting results were obtained by researchers aiming to distinguish between two or moretypes of cells (e.g., diseased versus normal, or cells with different types of cancers), based ongene expression data in the case of DNA microarrays. Since microarray data have large amountof data and attributes, which makes complex for researcher to do analysis. A small subset ofgenes is easier to analyze as opposed to the set of genes available in DNA microarray chips.Therefore it is important to focus on very few genes to give insight into the class association fora microarray sample. It also makes it relatively easier to deduce biological relationships amongthem as well as to study their interactions.In paper **(Shah, Marchand, & Corbeil, 2012)**they obtained feature selection algorithms for classification with tight realizableguarantees on their generalization error. The proposed approaches are a step toward which aremore general learning strategies that combine feature selection with the classification algorithm.and have tight realizable guarantees. They classified microarray data, where the attributes of thedata sample correspond to the expression level measurements of various genes was considered.They chosen decision stumps as learning bias, which is in part been motivated by thisapplication.(Banerjee, Mitra, Member, & Banka, 2007).In this paper they introduced an evolutionary rough feature selection algorithm for classifyingmicroarray gene expression pattern. Microarray data typically consist of large number ofredundant features; therefore an initial redundancy reduction of attributes was done to enablefaster convergence. The main aim was to retain only those genes that play a vital role indiscerning between objects. Rough set theory was employed to generate reducts, which represent the minimal sets of non redundant features capable of discerning between all objects, in amultiobjective framework.

V EXPERIMENTAL ANALYSIS

Lung cancer dataset was used to compare different filter based feature selection methods for the prediction of disease risks. Four classification algorithms reviewed above were considered to evaluate classification accuracy. The feature selection methods are

CSEBT-CfsSubsetEval_BestFirst
CSEGS- CfsSubsetEval_GeneticSearch
CLSEBFDT- ClassifierSubsetEval_BestFirst_ Decision Tree
GS- Greedy_stepwise
GSDT- GreedyStepwise_ Decision Tree
PCA- Principal Component Analysis
TRF- Tree RandomForet
TSC-Tree Simple Cart
TJ48-Tree J48
BBN-Bayes.BayesNet
BNB- Bayes.NaiveBayes
FRBFN-Function.RBFNetwork
FMLP-Function.MultilayerPerceptron

At first, feature selection methods were used to find relevant features in the lung cancer dataset and then, classification algorithms were applied to the selected features to evaluate the algorithms. Same experiment was repeated for four classifiers. WEKA 3.6.8 software was used. WEKA is a collection of machine learning algorithms for data mining tasks and is an open source software. The software contains tools for data pre-processing, feature selection, classification, clustering, association rules and visualization. ome performance measures were used for the evaluation of the classification results, where TP/TN is the number of True Positives/Negatives instances, FP/FN is the number of False Positives/Negatives instances. Precision is a proportion of predicted positives which are actual positive:

The following table show the experimental result of gene expression data set . Result show performance of various Attribute selection mode.

Cancer Data Set
Name:- Brain Tumour (Malignant glioma types)
Instances: 50
Attributes: 10368

Table 1
Attribute selection Performance

Sr. No	Evaluation Algorithm	Evaluator	Parameters Tuning	Atrribute Selection Mode	Evaluation mode
1	Attribute Subset Evaluator	CFS Subset Evaluator	Best first Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 764460 Merit of best subset found: 0.996	Including locally predictive attributes	Evaluate on all training data
			Greedy Stepwise (forwards) Start set: no attributes Search direction: forward Merit of best subset found: 0.996	Including locally predictive attributes	Evaluate on all training data
			Genetic search Start set: no attributes Population size: 20 Number of generations: 20 Probability of crossover: 0.6 Probability of mutation: 0.033 Report frequency: 20 Random number seed: 1	Including locally predictive attributes	Evaluate on all training data
			Linear Forward Selection Start set: no attributes Forward selection method: forward selection Stale search after 5 node expansions Linear Forward Selection Type: fixed-set Number of top-ranked attributes that are used: 50 Total number of subsets evaluated: 11148 Merit of best subset found: 0.968	Including locally predictive attributes	Evaluate on all training data
	Attribute Subset Evaluator	Classifier Subset Evaluator	Best first Classifier-ZeroR Start set: no attributes Search direction: forward Stale search after 5 node expansions Total number of subsets evaluated: 82914 Merit of best subset found: 152.942	Including locally predictive attributes	Evaluate on all training data
		Classifier Subset Evaluator	Genetic search Classifier-ZeroR Start set: no attributes Population size: 20 Number of generations: 20 Probability of crossover: 0.6 Probability of mutation: 0.033 Report frequency: 20 Random number seed: 1	Including locally predictive attributes	Evaluate on all training data

Sr. No	**Algorithm**	**FST**	**Total Number of Features Brain Tumour (Malignant glioma types)**	**Selected Features**
1	CFS Subset Evaluator	Best first	**10368**	99
2	CFS Subset Evaluator	Greedy Stepwise	**10368**	95
3	CFS Subset Evaluator	Genetic search	**10368**	4148
4	CFS Subset Evaluator	Linear Forward Selection	**10368**	39
5	Classifier Subset Evaluator	Best first/Decision Table	**10368**	04
6	Classifier Subset Evaluator	Genetic search	**10368**	1484

VI RESULTS

Cancer dataset was used to compare different feature selection methods for the prediction of disease risks. Six feature selection techniques are usedwith classification algorithms. CFS Subset Evaluator with Genetic search is performed better result as compare to other feature selection algorithm.

VII CONCLUSION

This feature selection algorithms shows that the feature selection algorithmconsistently improves the accuracy of the classifier. Each feature selection methodology has itsown advantages and disadvantages. Each algorithm has different behavior which shows thatusing single algorithm for different dataset is infeasible. The feature selection algorithms are onewhich decides the accuracy of the classification of different datasets. The feature selectionalgorithm must select the relevant features and also remove the irrelevant and inconsistentfeatures which cause the degradation of accuracy of the classification algorithms. Featureselection algorithm is playing a major role in accurate classification of large data set like geneexpression. Therefore proper cancer classification can be achieved using feature selectionalgorithms, and on time and accurate treatment may be provided to the patients.

REFERENCES

[1] Banerjee, M., Mitra, S., Member, S., & Banka, H. (2007). Evolutionary Rough Feature Selection in Gene Expression Data, 37(4), 622–632.

[2] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

[3] Langley, P. A. T., Flamingo, L., & Edu, S. (1994). Selection of Relevant Features in Machine Learning, 127–131.

[4] Shah, M., Marchand, M., & Corbeil, J. (2012). Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data, 34(1), 174–186.

[5] World health organization (2018). Cancer (https://www.who.int/en/news-room/fact-sheets/detail/cancer)

[6] Zhu, S., Wang, D., Yu, K., Li, T., & Gong, Y. (2010). Feature Selection for Gene Expression Using Model-Based Entropy, 7(1), 25–36.