IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PRE-PROCESSING OF HIGH DIMENSIONAL GENE EXRESSION DATA WITH GENE FILTERING IN CANCER DETECTION

[1]E.Monica Sushil Cynthia,[2] S.Kannan,

[1]Research Scholar,[2]Associate Professor,

[12]Department of Computer Application, School Of Information Technology

[12]Madurai Kamaraj University, Madurai, Tamil Nadu, India.

Abstract: In cancer diagnosis and drug creation, the recognition of various tumor types is important. The majority of previous cancer classification studies, on the other hand, has been clinically based and has limited diagnostic potential. The keys to solving the fundamental problems related to cancer detection and drug development are known to be found in cancer classification using gene expression data. Researchers have begun to investigate the possibilities of cancer classification using gene expression data as a result of the abundance of gene expression data. Because of the dataset's scale, visual analysis cannot be used to determine the number of missing values. This paper depicts the novel approach of pre-processing phases of gene expression data using gene pre-processor, with a focus on the filtering and normalization steps that can effectively deal with missing values in a dataset with thousands of rows and columns for cancer type classification. The experimental analysis shows that the proposed approach of gene pre-processor produces efficient results having reduced the size of the dataset with less elapsed time.

Index Terms **– Pre-processing, High Dimensional data, Gene Expression, Gene-Preprocessing.**

I. INTRODUCTION

Cancer research is one of the most important fields of medical research. The ability to reliably predict different tumour forms is extremely useful in terms of delivering better care and reducing patient toxicity. Previously, cancer was graded primarily on the basis of anatomy and clinical results. These traditional cancer classification approaches are said to have a number of diagnostic weaknesses [2]. Specifications of therapies focused on tumour forms differentiated by pathogenetic patterns have been suggested to enhance patient efficacy [3-11]. Furthermore, the current tumour groups have been discovered to be heterogeneous, consisting of diseases that are molecularly distinct and have varying clinical outcomes.

Systematic methods focused on global gene expression analysis have been suggested to obtain a deeper understanding of the issue of cancer classification. Gene expression levels are believed to hold the keys to solving fundamental problems in disease prevention and cure, biological evolution processes, and drug discovery. The recent advancement in cancer classification using gene expression data [4,12-15] was prompted by the simultaneous monitoring of thousands of genes enabled by microarray technology. Despite the fact that it is still in its early stages of growth, the findings obtained thus far appear encouraging.

Different statistical and machine learning classification approaches have been applied to cancer classification, but there are some problems that make it a difficult task. The gene expression data is unlike any other data that these approaches have encountered before. To begin with, it has a high dimensionality, with thousands to tens of thousands of genes. Second, the amount of publicly accessible data is extremely limited, with all of it falling below 100. Third, the majority of genes have no bearing on cancer classification. Established classification methods were obviously not designed to manage this form of data efficiently or effectively. Some researchers suggested that cancer classification be followed by gene selection. Gene selection reduces the size of the data, which improves the running time. Gene selection, meanwhile, eliminates a significant number of irrelevant genes, improving classification accuracy [1, 16].

Gene selection is a useful pre-processing technique in data mining that is typically used to minimise data dimensions and increase classification accuracy [17,18]. Using gene pre-processor, a novel approach to gene selection methods on genomic data has been proposed [16,19,20]. Firstly, the gene expression data set has very distinct characteristics that set it apart from all previous classification data. The bulk of publicly accessible gene expression data exhibits the following characteristics:

• high dimensionality: up to tens of thousands of genes,

• very small data set size: less than 100, and most genes are not related to cancer classification.

The pre-processing phases of gene expression data and emphases mainly on the filtering and normalization steps as the choices made here importantly influences the set of probe will be utilized in later investigations. In the filtering phase, two parameters are set. Initially, a recognition p-value cut-off for each analysis is determined, and the presence of a given sample is determined if its identification p-value is less than this cut-off. Second, a current limit is defined. It is used to determine how many samples a review

must appear in before being included in the dataset. After filtering, the information can be normalized. Dissimilar methods are assessed, and for the accessible dataset, normalization of the information on original scale gives the most stable outcomes.

Microarray analysis is used to monitor the expression patterns of tens of thousands genes simultaneously. Gene expression data can be presented as a matrix, with each row corresponding to a gene and each column representing a specified condition .The specific conditions usually relate to environments, cancer types or subtypes. Each entry is numeric representation of the gene expression level under a given condition with respect to a particular gene and tissues .Data mining is mainly concerned with the automated extraction of hidden predictive information from (large) databases. Preprocessing and post processing steps are often the most critical elements determining the effectiveness of real-life data-mining. From the point of medical sciences, data mining is involved in discovering various sort of metabolic syndromes. However, early detection and proper care management can make a difference in the health and longevity of individual sufferers.

1.1 Motivation of the Proposed Approach

The purpose of proposing Gene Pre-processor is to find the gene expressions from the huge datasets of high dimensional quality is motivated by the following points:

- Lack of identifying which genes should be selected from microarray profiles.
- How to select the minimum number of these genes sufficient for good diagnostic for classification of cancer associated gene.
- Less accuracy occurred in some existing system that had suggested statistical procedures. But there was no consensus about these procedures.
- Lack of increases of confidence and validity of the selected gene using existing algorithms.

These are the points from which the author got motivated and the Gene Pre-processor is proposed in which high quality data are considered with log transform expression ratio<9. This system proves to be more efficient in finding the type of cancers accurately.

1.2 Dataset

The datasets are downloaded from Gene Expression Omnibus (GEO) repository of NCBI. The GEO accession number of prostate cancer: GDS5072 (having 30331 number of genes over 11 sample condition); breast cancer: GDS5076 (having 13291 number of genes over 4 sample condition) lung cancer: GDS5040 (having 33297 number of genes over 6 sample condition). The datasets are freely available online in the soft file format.

1.3 Structure of the Paper

Section-II deals with the problems identified and Section III deals with the pre-processing preliminary steps. The proposed methodology which overcomes the above stated problems are discussed in Section IV. The Experimental results are discussed in Section V. The paper concludes with Section VI.

II. PROBLEM STATEMENT

- The size of the publicly accessible gene expression data set is still limited. The samples can be represented as very sparse points in a very high dimensional space if they are mapped to points in the attribute space. The majority of current classification algorithms were not created with this type of data in mind. Most classification algorithms face a significant challenge in such a sparse and high-dimensional situation.
- Overfitting is a major issue due to the large dimension, which is compounded by the limited data size. Furthermore, with so many genes in the tuple, computation time would be a significant challenge. As a result, designing an accurate and efficient cancer classification algorithm is a difficult challenge.
- The presence of noise in the data set raises the problem. Biological noise and scientific noise [22] are the two types of noise. Biological noise refers to the noise introduced by genes that aren't essential for classifying cancers. In reality, the majority of the genes have little to do with cancer types. Technical noise is associated with the non-uniform genetic histories of the samples or the misclassification of the samples, while biological noise is associated with the noises introduced at different stages of data preparation. The presence of noise, when combined with a limited sample size, makes accurate data classification difficult.
- Dealing with a large number of meaningless attributes is the biggest challenge (genes). Though irrelevant attributes can be found in almost all types of data sets studied previously, the ratio of irrelevant to important attributes in gene expression data is much lower. The existence of these unrelated genes reduces the ability of those relevant attributes to discriminate. This not only adds additional computation time to the classifier's training and testing phases, but it also raises the classification difficulty. Incorporating a gene selection system to select a group of gene expression is one way to deal with this. The above problems have been overcome by the proposed method of Gene Pre-processor whose methodology is dealt with in section III.

III. PRE-PROCESSING PRILIMINARY STEPS

The following subsections deals with the pre-processing preliminary steps to be kept in mind while cleaning the data acquired from the large dataset.

3.1 Dataset Description

The gene expression pattern datasets are then large tables with thousands of rows relating to the genes or clones in the DNA array and multiple columns, one for each estimated experimental disorder. A popular option for pre-processing gene expression patterns is to use a spread sheet, but even simple calculations can be inconvenient in a limited workstation due to the amount of memory requisite.

Because of the large dataset, visual analysis cannot be used to determine the number of missing values, search whether any gene has too many misplaced values, or look for duplicate genes in the DNA array. Doing these operations on a web server, on the other hand, allows you to take advantage of the server's capabilities while still ensuring that you're using the most recent version of the program. Finally, using a common interface to parse the input eliminates potential problems caused by complex file formats, resulting in a clean and standard file that can be analysed with various tools.

They discovered that as the information is modified and slightly different choices are made in the pre-processing stages, the subsets of genes that are selected vary for different versions of the dataset. To fully understand the importance of the choices made during the pre-processing stages, we looked at what choices are being made and how each of these affects the final collection of genes. An information matrix of gene expression values in blood cells, a data matrix with negative controls, and context information, which includes clinical and technical data, make up the dataset. Every case sample is coordinated to its consistent control through the case/control identifier given in the background information.

3.2 Information Cleaning Process

The main stages in the information cleaning process are:

- Remove probes for genes required with gene expression data.
- Remove case-control pairs when either the case or control is an outlier with respect to quality, and expel the pairs.
- Perform background correction using negative control probes.
- Filter out probes that are not perceptible or adequately present and translate from probes to genes.

3.3 Relative Expression Levels

The array is agitated by a laser to determine the relative abundance of the hybridised RNA. The location would be red if the RNA from the sample population is abundant, and green if the RNA from the control population is abundant. The spot will be yellow if the sample and control bind equally, but it will not fluoresce and appear black if neither binds.As a result, the relative expression levels of the genes in the sample and control populations can be determined based on the fluorescence intensities and colours for each spot.

3.4 Gene Expression and Microarray Technology

DNA serves as a template for creating copies of itself, as well as a blueprint for the RNA molecule (ribonucleic acid). The genome serves as a blueprint for the development of a wide range of RNA molecules. Messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA are the three major types of RNA (rRNA).The genetic information contained in the DNA molecule is distributed in two stages: The DNA molecule is transcribed into mRNA in the transcription stage, and mRNA is translated into the amino acid sequences of proteins that perform different cellular functions in the translation stage.

Gene expression is the method of translating a gene's DNA sequence into RNA. The expression level of a gene refers to the number of copies of that gene's RNA produced in a cell and is inversely proportional to the amount of corresponding proteins produced. Relevant patterns of gene expression have been observed during various biological states, including embryogenesis, cell growth, and normal physiological responses in tissues and cells [21].As a consequence, a gene's expression provides a measure of its function under particular biochemical conditions. It is well understood that many diseases, such as cancer, are expressed in changes in gene expression values. A series of mutations in genes that regulate the cell cycle, apoptosis, and genome integrity, among other things, can turn normal cells into malignant cancer cells [22].

We can create gene expression profiles that depict the complex functioning of any gene in the genome by estimating transcription levels of genes in an organism under various conditions, at different developmental stages, and in different tissues.We may visualise the expression data in a matrix with rows displaying genes, columns displaying samples (e.g., various tissues, developmental stages, and treatments), and each cell containing a number indicating the level of expression of a specific gene in a specific sample. A gene expression matrix is the term for such a table.Creating a database of these matrices will help us better understand gene regulation, metabolic and signalling pathways, disease genetic processes, and drug response. For example, if overexpression of specific genes is related to a specific cancer, we should look at what other factors affect their expression and which other genes have similar expression patterns

Microarrays and serial gene expression analysis are two new methods for simultaneously calculating thousands of genome-wide expression values. Microarrays indirectly measure the target quantity (i.e. relative or absolute mRNA abundance) by measuring the intensity of the fluorescence of the spots on the array for each fluorescent dye, i.e. for each optical wavelength. As a result, the raw data produced by microarrays is in fact monochrome images. The method of converting these images into a gene expression matrix is not simple: the spots that correspond to genes on the microarray must be identified, their boundaries determined, and the fluorescence intensity from each spot measured and correlated with the background intensity and other channel intensities.

Microarray-based gene expression measurements are also a long way from supplying counts of mRNA per cell in a sample. The measurements are comparative in nature: we can compare the expression levels of similar genes in dissimilar samples with the expression levels of dissimilar genes in the same sample. All knowledge contrasts must also be subjected to sufficient normalisation. The ratios' accuracy is based on absolute intensity values, which vary from spot to spot due to sequence specificity and cross-hybridization of homologous sequences.

When looking at the gene expression matrix, this must be held in mind. Consistency and restrictions of particular microarray systems for specific types of measurements, as well as cross-platform association and normalisation, will greatly increase the utility of microarray-based gene expression measurements.

IV. Proposed Methodology

The preprocessing functions incorporate:

4.1 Transformation

After we've entered the raw image data into the gene expression matrix, the next step is to analyse it to see whether we can extract some information about important biological processes. There are two basic approaches for analysing the gene expression matrix:

1. Connecting rows in the expression matrix to connect expression profiles of genes;
2. Connecting columns in the expression matrix to connect expression profiles of samples.

We should look for similitudes or dissimilarities when associating rows or columns. If two rows are found to be parallel, we can conclude that the genes in those rows are co-regulated and possibly functionally related. We will discover which genes are differentially expressed and, for example, research the effects of different compounds by associating samples.

Before we can perform any associations, we require an approach to measure the resemblance (or distance) between the objects we are associating. We can repute these objects (rows or columns in the matrix) as points in n-dimensional space or as n-dimensional vectors, where n is the number of samples for gene comparison, or number of genes for sample comparison. Normally microarray information are log transmuted. This is done for a number of reasons, containing: it stabilizes the variance; it compresses the variety of information; and it creates the information more usually dispersed, which permits statistics to be applied to the information.

- **Stabilizes the variance**: for information on the linear scale, as the expression level/signal gets higher, the variance also increases. Log transforming decreases this dependence, so that the variance is more reliable throughout the variety of expression.
- **Compresses the range of information**: In a typical microarray experiment, most genes have levels <100, with very few having large levels >10,000. When this information is log transmuted, the range shrinks from commonly 1-65,000 to 0-16.
- **Usually distributed information**: For performing statistics and the expectations is that the errors are usuallydispersed. This is undeniably not the case for unlogged information& this violation of the statistical assumption makes the statistical findings on unlogged data invalid.

4.2 Missing Value Removal

In a distinctive gene expression data matrix, the rows are the genes (or oligonucleotides) under examination and the columns are the experimental conditions or time points. The gene expression data matrix is attained by performing a sequence of microarray experiments on the similar set of genes, one for each column. Let the gene expression data be signified as an M N matrix Y where the entries of Y are the expression ratios for M genes under N dissimilar conditions or time points. Then the element y_{ij}signifies the expression level of the i^{th} gene in the j^{th} experiment. The target of missing value imputation is to guess the missing entries given the deficient gene expression data matrix Y.

Missing value imputation requires using knowledge about the data to determine the entries that have been lost. There are two types of knowledge that are commonly available. The correlation structure between entries in the data matrix is the first type of information. Since genes involved in parallel cellular processes have similar expression profiles, there is overlap between rows in a gene expression data matrix. Correlation occurs between columns as well, since the set of genes is assumed to behave similarly under similar conditions.

Following that, the missing entries could be evaluated based on a subset of associated genes or a subset of associated conditions. Domain knowledge about the data or the processes that generate the information is the second type of information. The domain information can be used to standard is the estimation process, resulting in more plausible estimates. More modern gene expression missing value imputation algorithms have endeavored to integrate information about the underlying biological processes to develop the imputation accuracy.

Missing values frequently happened due to numerous reasons, such as insufficient resolution, image corruption, dust or scratches on the slide and etc., though none dominated. Normally, microarray datasets are assessed to have more than 5% missing values and up to 90% of genes are influenced. As a result, although frequently ignored, the missing value imputation is necessary to minimalize the detrimental impact of missing values on the microarray information analysis.

Certainly, one methodology to validate the analysis technique of the microarray information with missing values is to recurrence the experiments, and observably it is very costly and time consuming. There are also numerous simple ways to handle missing values, e.g. eliminating the genes with missing values from additional analysis, substituting missing values by zeros, or filling the missing values with the row or column averages accessible. These methods are not ideal because they did not deliberate the correlation of the information, which empowered the advancement of more refined missing value procedures that attempted to exploit the information relationships by utilizing the information accessible in the entire dataset.

In the context of gene expression data, MV imputation techniques typically fall into two types. In the first type ("local" approaches), the expression information of a missing entry is taken from neighbouring genes, where their nearness is determined by a proximity measure (e.g., correlation, or the Euclidean distance). For the second type ("global" approaches), dimension reduction methods are applied to decay the data matrix and iteratively rebuild the missing entries.

Then this analysed the impacts of these imputation strategies, as well as the Mean and Median methods, where MVs are exchanged with a simple mean or median, correspondingly, from identified values. Given these approaches we conducted two classical downstream analyses for cancer gene expression data: classification and clustering. We utilized the following experimental design to do so. First, we detached all genes with more than 10 % missing values (MV filtering). Next, we imputed the missing values utilizing each of the five approaches. Finally, we applied a non-supervised filter to eliminate genes with slight variation between samples. Additionally, this value drops to an average of 2.32% after the MV filtering. This can be perceived as indicative of the upper bound values for experimental settings with artificially imputed missing values. This designates a minimal influence of the imputation technique on the non-supervised filtering step.

4.3 Filtering Data

This method for gene filtering involves removing genes belonging to components, which we expect to contain mostly non-informative genes. It is known that genes corresponding to either low values of mean expression or to low values of variance of expression are more likely to be non-informative. The same property should pertain to Gaussian components. When we decompose the sample means or sample variances into Gaussian components, we can order the components with respect to their location parameter (mean of the Gaussian component). Then we remove genes which belong to components located at the left hand side of the signal scale, i.e., with the lowest values of this parameter. We assume that their inclusion into the further analysis would lead rather to false discoveries than to detection of true differentially expressed genes (DEGs).

The issue is how many components consistent to low values of x should be detached. We suggest and examine two approaches for selecting the number of components to eliminate. The first one is based on the "top three" rule. More precisely, we expect that three components with maximum values for parameter of location, named high-level expressed genes, medium-level expressed genes and low-level expressed genes, are informative and we preserve genes consistent to these components. Other genes are detached. The second technique is to utilize a clustering process, which categorizes assessed Gaussian components into two groups.

Filter Flat Patterns (no pertinent dissimilarities between classes):

- By number of peaks
- By root mean square
- By standard deviation

-Standardize Patterns: Subtracts the mean of the pattern and split it by the standard deviation (z-score).
The main motivation for gene filtering:

- Comparatively few genes must be: expressed at any time
- Comparatively few genes must be: differentially expressed between conditions
- Restrict attention to those genes that are:ν relevant "candidates"

This system has selected k-means clustering in three dimensional space with coordinates given by means, standard deviations and weights of Gaussian components. The K-means algorithm reduces the within-cluster sums of squared Euclidean distances from every point to the centre of the cluster. The number of clusters is presumed equal to 2. Two three dimensional clusters are well-ordered with respect to their location along the "mean of Gaussian component" coordinate. Then the cluster which location along this coordinate relates to a lesser value is deliberated non-informative. Accordingly, genes that belong to the Gaussian components within this cluster are expelled.

Filtering strategies can be utilized to diminish the number of tests and subsequently increase the power to identify true differences. An ideal filtering technique would eliminate tests which are actually null (consistent to genes that are correspondingly expressed), while leaving those tests consistent to genes which are actually differentially expressed. Numerous approaches for filtering have been proposed comprising filtering by variance, signal, and MAS detection call.

In quality filtering phase all bad quality information is expelled. This incorporates genes with absent or immersed measurements. In information quality matrix first column demonstrates if sample is control (positive or negative), next two incorporates information about immersion for r and g and final two designate if surrogates (replacement for missing measurement) are utilized. Though microarray dataset comprises a huge number of genes, a part of genes are usually excepted during the expression profiling. This procedure, i.e. *Gene Filtering*, is targeted at eliminating the unwanted-genes that comprise outliers and too much misplaced expression values, and that do not display variability across tissue samples.

Consequently, we select a filtering technique that discards numerous unreliable and uninformative data points, whereas accepting the majority of gene expression. Once more is identified about the variation of expression measurements, it will be probable to design filters that differentiate actual variations in expression level from background noise and measurement error. In the nonexistence of such information, numerous dissimilar filters must be tried and subjective filtering parameters must be selected on the basis of the particular data set.

Once the data set has been filtered, we discover that it is beneficial to scale the expression level of each gene to have mean zero and variance one. This catches the notion that the expression patterns of two genes might be parallel in shape, even though one is conveyed at a much greater level than the other. We will denote to a data set that has been scaled in this manner as the consistent information.

4.4 Gene Pre-Processor

The proportion between the sizes of the detached and retained gene pools is an important parameter to choose. We propose a new approach for evaluating near-optimal threshold values for sample means and sample variances for gene filtering.We show that our adaptive technique improves the sensitivity of finding differentially expressed genes as compared to previous techniques of filtering microarray data using fixed threshold values by analysing a large number of publicly available datasets and simulated datasets.

The technique for gene filtering entails removing genes from elements, which we believe to be mostly non-informative genes. It is well known that genes with low mean expression or low variance of expression are more likely to be non-informative. Gaussian components must have the same property. We may order the components with respect to their position parameter by decaying the sample means or variances into Gaussian components (mean of the Gaussian component). The genes that belong to components on the left hand side of the signal scale, i.e. those with the lowest values of this parameter, are then eliminated. We anticipate that incorporating them into the investigation would result in more false disclosures than real DEG identification.

The issue is how many components relating to low values of x should be evacuated, we suggest “Top three” model. More precisely, we expect that three components with maximum values for parameter of location, named high-level expressed genes, medium-level expressed genes and low-level expressed genes, are informative and we recollect genes consistent to these components. Other genes are detached. The pseudo code for the proposed approach is as follows:

Input: S-Size of the gene
X-Column Matrix
L-Length of the gene
Output: P-Preprocessed Data

1: Read gene set from the input dataset
2: Identify the size of the selected gene set
3: Represent the gene set to a vector
4: Compute the length of the gene for the selected set
5: Then applying gene filtering condition for the number of genes in the dataset
6: Obtain every gene from the set and check for the missing values
7: If the missing value > the select gene
8: Then check for over and under expression genes with the condition Max(D) and Min(D)
9: After that select genes top N of the highest maximum values
10: Finally represent the filtered gene data

End Algorithm

V. EXPERIMENTAL ANALYSIS

5.1 Dataset Reduction

The data of gene expression are collected from the huge dataset as mentioned in Section I. It is a huge dataset that contains null values, missed values etc. which has to be pre-processed efficiently by using an efficient pre-processor. Only then the needed information can be processed effectively. A sample of two soft files has been shown in Fig.1 which shows the reduction of datasets from the huge database by eliminating the unwanted information from the database after the pre-processing stage.

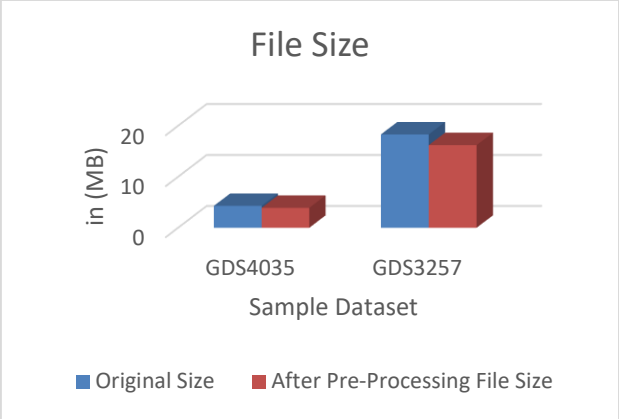

Fig.1 shows the reduction of file size post pre-processing

Table.1 shows the amount of data that have been eliminated from the huge dataset for a sample of only two soft files. The amount of file size is shown in MB.

ID	Dataset	Original Size	After Pre-Processing File Size
1	GDS5072.soft	4.29 MB	3.9 MB
2	GDS5076.soft	18.3 MB	16.2 MB

Table1. Exploration of the dataset reduction in MB

5.2. Elapsed Time

Elapsed time can be defined as the actual time taken to complete a process. As mentioned in subsection 5.1, the dataset reduction from huge database takes some amount of time for the process to be done. Hence the elapsed time for the dataset reduction of our proposed approach is less. Fig.2 shows the elapsed time of the dataset reduction and table.2 shows the elapsed time in seconds for a sample of only two soft files.

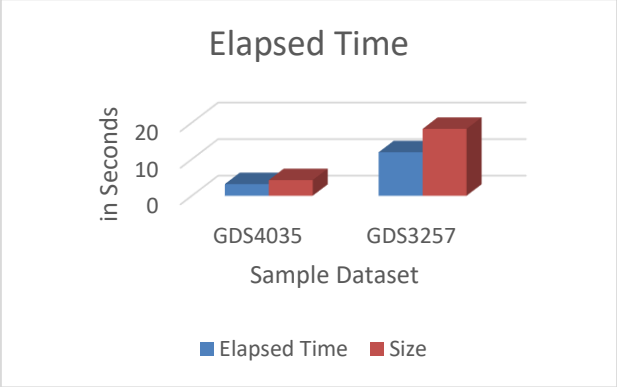

Fig.2 show the elapsed time in seconds

ID	Dataset	ET (Elapsed Time)
1	GDS5072.soft	3.2 Sec
2	GDS5076.soft	11.9 Sec

Table.2 Numerical values of the Elapsed Time for the soft files in seconds.

6. Conclusion

Cancer research and drug discovery benefit greatly from a systematic and impartial approach to cancer classification. Previous cancer classification techniques were all clinically based and had diagnostic limitations. Gene expression has long been thought to hold the key to solving the basic problems of cancer diagnosis, treatment, and drug development. Microarray technology has recently made it possible to produce large quantities of gene expression data. This has prompted scientists to propose various cancer classification algorithms based on gene expression results.

In this paper, Gene Pre-processor has been proposed for the pre-processing tasks. Each pre-processing task had been done so efficiently that no value is missed, noise has been eliminated etc. The experimental analysis shows that our proposed Gene Pre-processor is more efficient in pre-processing the gene expressions achieving huge reduction of file size from a huge database with

less elapsed time. This huge reduction of file size with less elapsed time can produce a high accuracy rate in further processing stages.

VI.ACKNOWLEDGMENT

REFERENCES

[1]. Ying Lu, Jiawei Han. 2003. "Cancer Classification Using Gene Expression Data", Information Systems. 28(4): 243-268.
[2]. A. Azuaje.2000. "Interpretation of genome expression patterns: computational challenges and opportunities", IEEE Engineering in Medicine and Biology.
[3]. A. Alizadeh. 2000. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. Nature. 403:503–511.
[4]. T.R. Golub, D.K. Slonim, P. Tamayo, M. Gaasenbeek C. Huard, J.P. Mesirov, H. Coller, M. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science.531–537,
[5]. L. Veer, H. Da, M. Bijver. 2002. Gene expression profiling predicts cinical outcome of breast cancer. Nature.530–536.
[6]. T. Sorliel.2001. Gene expression patterns of breast carcinomas distinguish tumor subclass with clinical implications. In Proc. of National Academy of Science.10869–10874.
[7]. S. Pomeroy, P. Tamayo, M. Gassenbeek, and et al. Prediction of central nervous embryonal tumour outcome based on gene expression. Nature, pages 436–442, 2002.
[8]. T. Sorlie and et al. 2001. Gene expression patterns of breast carcinomas distinguish tumor subclass with clinical implications. In Proc. of National Academy of Science. 10869–10874.
[9]. D. Zajchowski and et al. 2001. Identification of gene expression profiles that predict the aggressive behavior of breast cancer cells. Cancer Research. 5168–5178.
[10]. W Dubitzky, M Granzow, and D Berrar. 2002. Comparing Symbolic and Subsymbolic Machine Learning Approaches to Classification of Cancer and Gene Indentification. Kluwer Academic.
[11]. L. Veer and D. Jone. 2002. The microarray way to tailored cancer treatment. Nature Medicine, pages 13–14, Jan 2002.
[12]. J. DeRisi, L. Penland, P. Brown, and et al 1996. Use of a cdna microarray to analyse gene expression patterns in human cancer. Natural Genetics, 4:457–460.
[13]. D. Slonim, P. Tamayo, J. Mesirov, T. Golub, and E. Lander 2000. Class prediction and discovery using gene expression data. In Proc. 4th Int. Conf. on Computational Molecular Biology (RECOMB). 263–272.
[14]. S. Lakhani and A. Ashworth 2001. Microarray and histopathological analysis of tumours: the future the past? Nature Reviews Cancer, pages 151–157.
[15]. D. Nguyen and D. Rocke 2002. Classification of Acute Leukemia based on DNA Microarray Gene Expressions using Partial Least Squares. Kluwer Academic.
[16]. A. Berns.2000.Cancer: Gene expression in diagnosis. Nature, pages 491–492.
[17]. I. Guyon, J. Weston, S. Barnhill, M. D., and V. Vapnik. 2000. Gene selection for cancer classification using support vector machines. Machine Learning.
[18]. D. Koller and M. Sahami.1996. Towards optimal feature selection. In Machine Learning: Proc. of 13th Int. Conf.
[19]. W. Siedlecki and Sklansky 1998. On automatic feature selection. International Journal of Pattern Recognition and Artificial Intelligence, 2:197–220.
[20]. E. Xing, M. Jordan, and R. Karp.2001. Feature selection for high-dimensional genomic microarray data. In Proc. of the 18th Int. Conf. on Machine Learning.
[21].C. Campbell, Y. Li, and M. Tipping.2001. An efficient feature selection algorithm for classification of gene expression data.
[22]. P. Russel. Fundamentals of Genetics. Addison Wesly Longman Inc., 2000
[23]. A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classication with gene expression profiles. In Proc. of the Fourth Annual Int. Conf. on Computational Molecular Biology.
[24]. M. Schena, D. Shalon, R. Davi, and P. Brown.2000.Quantitative monitoring of gene expression patterns with a complementary and microarray. Science, 270:467–470, 1995.
[25] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, and H. Horton. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology, 14:1675–1680.

VI.ACKNOWLEDGEMENT

E.Monica SuShil Cynthia, Research Scholar, Department of Computer applications,Madurai Kamaraj University, Madurai. She completed her UG Degree (B.Sc (Physics)) LDC, Madurai in 1998, and PG Degree (M.C.A) in American College, Madurai in 2000. She has 15+ years of experience in teaching field. She has published 5+ research papers in journals and conferences. She has interests in domains like, Image Processing, Data Mining and Networks. Email: monikasushil777@gmail.com.

Dr.S.Kannan, Professor in Computer Application, Madurai kamaraj University, Madurai.He has 24+ years of experience in teaching field and 12+ experience in research field. He has published 20+ research papers in journals and conferences. He has interests in domains like Data Mining, Image Processing and Networks, Data structures and algorithm, Network Security and Soft Computing. Email: skannanmku@gmail.com

.