CLASSIFICATION OF CARCINOMA GENES BY USING HYBRID METHOD OF HIDDEN MARKOV MODEL & ADAPTIVE NEURO FUZZY INFERENCE SYSTEM

Dr. M. Sangeetha

Professor & Head, CSE Department, V.S.B. Engineering College, Karur.

ABSTRACT—Process of knowledge discovery in databases is an important step in Data mining, in which intelligent methods are applied in order to extract patterns. In recent time medical researchers are used data mining techniques for study of disease research, especially most dangerous disease like cancer. Generally cancer research is clinical in nature. Classifying and predicting the outcome of a disease is a significant and demanding tasks in data mining. The medical research group contains collected and made available large volumes of medical data. Hence, medical researchers are used popular Knowledge Discovery Databases (KDD) research tool for identify and exploit patterns and large number of variables relationships, and also prepared to enable these medical data to predict the disease outcome using the historical cases stored within datasets, which also includes data mining technique. Here this proposed work provides an overview of the current research being carried out on Leukemia cancer dataset using data mining techniques that is classification technique to enhance cancer diagnosis. Although cancer classification has improved over the past few years in data mining, there is no general approach for identifying new cancer classes (class discovery), in order to overcome this problem here a most generic approach to cancer classification method that is K-Nearest Neighbour (KNN) classification with Geometric Particle Swarm optimisation (GPSO) feature selection based on gene expression monitoring by DNA microarrays is described and applied to human leukemia dataset as a test case.

Keywords-- Data mining, Knowledge Discovery in Databases (KDD), K-Nearest Neighbour (KNN) classification with Geometric Particle Swarm optimisation (GPSO)

1. Introduction

Microarray Technology[1] is one of the significance method in investigating the enormous amount of genes as the changes in the microbes are related to the changes in the gene patterns. Different gene expression profiles are compared from tumors such as Leukemia [2], Colon [3] and Breast [4] and the tissue expressions are examined and compared. These data are very much unessential and may contain unwanted data and most of the genes provide not useful information. As a result, the most important thing is the tools to deal with these issues in efficacy. It is essential to categorize a set of necessary genes from a high volume data set polluted with maximum noise [5].

To examine these data, Feature selection is the best process to reduce the capacity of data set and to improve the analysis process [6]. In Gene expression analysis, Feature Selection uses a number of categorization methods [7] to classify a category of tumor, and to minimize the count of genes to investigate in case of a current patient, so that detecting and diagnosing has been done at the earlier stage. Different types of classification techniques such K-Nearest Neighbor (K-NN) [8] or else Support Vector Machines (SVM) [9] have been used. By applying data reduction on the number of considered genes, classification accuracy can be enahnaced. This system is proposed to discover features of genes and classify cancer using leukemia dataset. For that this proposed work has to use GPSO feature selection method and KNN classifier with pre-processing steps, which are commonly used in the field of data mining and pattern recognition. Feature selection method include Euclidean distance, and also, k-nearest neighbour (KNN) classifier shows the improvement in the performance of classification with use of distance parameter and feature selection method compare than the existing classification methods.

2. Literature Survey

Kai et al., [10] illustrated that a cancer categorization technique such as Discriminant Kernel-PLS has been used in the profile of gene expressions. In this work the dataset of prostate cancer, lung cancer and leukaemia are used. This method provided more prediction accuracy by using the 1-ANOVA.

Manuel *et al.,* [11] illustrated technique for gene expression profiles cancer categorization that was called as Kernel alignment-KNN. This categorizer has been used to retrieve the prominent results. This paper also explained about the Kernel alignment algorithm's linear combination of dissimilarities, these also performed well when it compared with other metric learning strategies and improved the classical k-NN based on a single dissimilarity.

Dimitris *et al.,* [12], detailed a medical diagnosis using gene expressions. Data pre-processing methods are utilized to find the missing value, then the feature selection methods are applied to reduce the dimensionality of the data set. Three datasets were also used. Finally the classification is done by using Novel SVM-based architecture. And final result has shown that this proposed gene expression analysis system for medical diagnosis provided 100% of result by three different types of dataset.

Wang *et al.,* [13] utilized a fuzzy based framework for medical diagnosis. In this work, author were used a fuzzy-based ensemble model and a widespread fuzzy-based framework for cancer classification of micro array gene expression data. This proposed fuzzy based system for micro array gene expression data is used both gene selection and classification methods to discover the data classes. This proposed Neuro-Fuzzy Ensemble model (NFE) makes fuzzy based system more realistic to gene profiles micro array. The performance of the proposed method attained by using fuzzy based system is more feasible.

Huang *et al.,* [14] provided a novel method for cancer diagnosis with use of discriminative genes from micro array gene expression data. New mutual information (MI)-based feature-selection resolved a micro array gene expression-based data large p and small n problem in more efficiently. And also this method to expressed reliable domino effect still when merely a small model suite is accessible. Additionally, novel MI-based criterion is planned to shun the very much superfluous choice, results in a methodical manner. Recently, through the evaluation of MI, the proper chosen trait subsets can be conceivably determined.

3. Dataset

This proposed methodology uses the instances from well known Leukemia Dataset for cancer cell gene classification experiment. Leukemia dataset was taken from the public GEMS Data Repository with url http://www.gems-system.org/.[15] The Leukemia dataset have two types such as Leukemia 1 and Leukemia 2. There are also different class of samples such as Acute myelogenous leukemia (AML), acute Lympboblastic Leukemia (ALL) B-cell, and ALL T-cell and mixed-lineage leukemia (MLL).

4. Proposed Methodology

Main aim of this proposed methodology for leukemia gene classification system is to design and develop an approach for the purpose of leukemia cancer cell diagnosis using gene classification method with feature selection process. There are several schemes that discover the cancer cell diagnosis in automatic way. The proposed Leukemia gene classification process is a better way of targeting a promising results with use of given data set based on data mining technique, here this proposed method have planned to make use of enhanced preprocessing Independent Component Analysis (EICA) method and Bi-PCA with Geometric Particle Swarm optimization (GPSO) based feature selection and classification. The proposed approach have to use the following steps for cancer and non-cancer cells classification process with use of leukemia data set, they are pre-processing, feature selection and classification methods. Early of this proposed method the given dataset is pre-processed for the purpose of efficient final classification result and missing value calculation and duplication elimination, which will be done with the assistance of the newly Enhanced Independent Component Analysis process. After that the analyzed dataset is used for feature selection process, in this stage the most important features are selected with the assistance of Geometric Particle Swarm Optimization (GPSO), then the selected features are used for classification process, this process will be done with the assistance of classifier. The proposed classifier, KNN will be utilized to classify the cancer and non-cancer cells based on the gene micro arrays. At last, the given dataset is subjected to the proposed technique to assess the performance in classifying the cancer or non-cancer cells from the leukemia dataset. In this proposed method, for experimentation, the dataset will be subjected to analyze the performance of the proposed approach utilizing Classification Accuracy, F-score Sensitivity and Specificity. Figure 1 shows the flow diagram of leukemia gene classification of using

the leukemia dataset with proposed methodology, which is shown below.

4.1 Preprocessing

The aim of the preprocessing stage in proposed leukemia gene classification process is eliminate the duplication values and calculate the missing values of the gene array leukemia dataset. In data discovery analysis process, preprocessing is one of the most important techniques of analyzing data prior to further computational processing. Here this proposed method used two methods Bi-PCA and Enhanced ICA, both methods are used to analyze the data for final leukemia gene classification process, in a maximum preservation of the original information of given dataset. This proposed Bi-PCA and EICA should perform superior to the existing schemes in terms of duplication elimination and missing value calculation of the dataset.

BiCluster Principal Component Analysis (Bi-PCA)

The given leukemia data set of gene expression profiles for cancer diagnosis is represented by a numerical $(M \times N)$ matrix X, where N is the number of genes and M is the number of samples. This proposed preprocessing method assumes biclusters to handle the local similarity structure of the matrix, where there is most interrelated rows and columns with the missing entry are chosen to estimate the missing value of the given dataset for leukemia gene classification. Bi-PCA technique is illustrated as below:

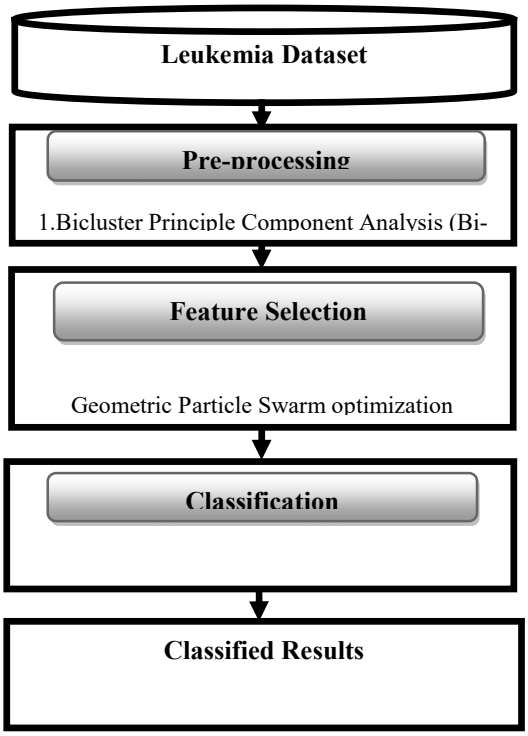

Figure 1. Proposed Leukemia Gene Classification Method

Initially bi-clusters are ordered clusters where rows contain correlated genes and columns contain correlated experimental conditions that indicate the missing value. The phenomenon of dual clustering was introduced early but it become popular since 2000 when Cheng and Church applied it in the gene expression matrices [16]. Gan *et al.* introduced a geometrical biclustering method [17], where biclusters embedded in a matrix can be regarded as points dispersed on special linear structures in high-dimensional space, and the Hough transform is used to identify these linear patterns in the high-dimensional space in order that biclusters can be identified. Other dual grouping methods are proposed based on distance measures [18], probability models [19], and hypergraph-based geometry [20].

PCA indicates the change of M −spatial gene vectors x as a precise combination of primary axis vectors $u_i (1 \leq i \leq K)$ whose count is small$(K < M)$;

$$x = \sum_{i=1}^{K} y_i u_i + \epsilon, (1)$$

The linear coefficients $y_i (1 \leq i \leq K)$ are called factor scores.

A specially dogged K number using in PCA, it obtains y_i and u_i, such that the sum of squared error $\|\epsilon\|^2$over the fully data set X is in minimum. When there is no missing value, y_i and u_i are calculated as follows.

A covariance matrix P for the expression vectors $i (1 \leq i \leq N)$ is given by

$$P = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T (2)$$

where μ =mean vector of x: $\mu \stackrel{def}{=} (1/N)\sum_{i=1}^{N} x_i$. T = the transpose of a vector or a matrix. Here also describe the $i-th$ principal axis vector by $u_i = \sqrt{\lambda_i} t_i$. With these notations, the i^{th} factor score for an expression vector x is given by $y_i = \left(\frac{u_i}{\lambda_i}\right)^T x$. Factor scores $y = (y1, \ldots, y_K)$ for the gene expression vector x obtained by minimize the residual error:

$$E = \|x^o - U_x^o\|^2 (3)$$

The least square solution of the missing value estimation problem is given by $= (U^{oT} U^o)^{-1} U^{oT} x^o$. Using y, the missing value is calculated as $x^{miss} = U^{miss} x$.

Here this work also finds a bicluster for every individual missing value. First, for every target gene, initial complete matrix is used to recognise by classification genes set in the first step, according to the euclidean distances between the target gene and all the other genes. In biclustering process, consider each individual condition has its own correlation with others. On the other side, i and j conditions $(1 \leq i, j \leq p, i \neq j)$ possess various correlations with other $n - p$ limitations. To take into account the correlation among the $n - p$ different conditions for the p missing entries in the target gene, introduce a matrix Q in (4) :

$$Q = C^T D (4)$$

The distances between the target gene and other genes for the jth missing entry are calculated by classification genes weighted Euclidean distance from the dataset (5):

$$dj(gt, gs) = \frac{\sqrt{\eta(v-p)^2[g_t(v) - g_s(v)]^2}}{\sqrt{\sum_{v=1}^{n-p} \eta(v)^2}} (5)$$

where $1 \leq j \leq p$ and g_t = the target gene, g_s = one of the other genes in the matrix, and $\eta(v)$ denotes the (j, v)th element of matrix Q. In (5), $\eta(v)$ serves as a weight for calculating the distance between g_t and g_s in the vth position. k smallest weighted euclidean distances are chosen in the corresponding genes which is to be the reselected neighbor genes for the jth missing value. The most correlated experimental conditions also have to be selected from the frequent correlated genes, this is decided by the value of $\eta_j(v)$. If $|\eta_j(v)| \geq T_0 \eta_j, max$, then the vth experimental condition is considered to be correlated with the jth missing value, where $\eta_j, max = max v \in \{1,2, \ldots n-p\} |\eta_j(v)|$, and T_0 is a preset threshold. After the uncorrelated genes and experimental conditions are removed, and get a subset D_j . The rows and columns of D_j are the most correlated genes and experimental terms and rules with the jth missing value of the objective gene, respectively.

The bicluster for the lth missing value is in this form:

$$bicluster_l = \begin{pmatrix} \alpha_j & w_j \\ c_j & D_j \end{pmatrix} (6)$$

where α_j denotes the jth missing value of the target gene, cj is the column in α_j 's position in the most correlated genes, w_j is the non-missing values in the most related locations with α_j , and D_j is the subset was find in prevoiusly. The missing value in a bicluster is α_j , target gene. At last this pre-processing work Conduct Bi-PCA for a second time on biclusters. For a target gene containing p missing values, conduct Bi-PCA in

$bicluster j$ $(1 \leq j \leq p)$ until the p missing values are estimated, and this work can get a complete gene vector for further processing in the proposed method.

Enhanced ICA

In this work another preprocessing technique EICA is used for analyzing the dataset for final classification process. Every IC component has genetic significance and relates to a specific gene signal. Here this work employed an enhanced ICA method which is shown in detail as below.

$$X(t) = B*V(t) \quad (7)$$

Where, a data matrix $X(t) = [X_1(t), X_2(t), \cdots, Xp(t)]^T$ with $p \times n$ dimensions, and its rows correspond with observed signals and its columns correspond with the number of samples. $B = [b_1, b_2, \cdots, b_m]$ is combination matrix with $p \times m$ dimensions and source signal matrix $V(t) = [V_1(t), V_2(t), \cdots, V_m(t)]^T$ with $m \times n$ dimen sions as its rows are independent statistically. Variables found in $S(t)$ rows are called ICs and $V(t)$ is the observed signals form a linear combination with these ICs. In other words ICs estimation is made with finding linear relation of observed signals.

By applying ICA, a combination of two matrixes B and $V(t)$, to achieved source signal. The i^{th} level of DNA microarray expression gene, x'_{ia} is reconstructed by i^{th} IC of IC_i (i = 1,⋯,p), according to relation (1) have:

$$x'_{ia} = b_i * V_i \quad (8)$$

Definitely, if gene expression level for main microarray's i^{th} gene is x'_{ia}, then error average square of renovated samples be:

$$E_i = \frac{1}{n}\sum_{j=1}^{n}|x_{ij} - x'_{ij}|^2, j = 1, \ldots, n \quad (9)$$

Then to calculate error average square amounts, and order these into reconstructed samples, and select p' IC components with lower error.

Feature Selection using Geometric Particle Swarm Optimization (GPSO)

Main objective of solving the gene selection problem, a novel PSO based algorithm discovered, based on the geometric framework in [22], which was developed in this proposed work. This novel method, namely known as Geometric Particle Swarm Optimization (GPSO), that enables to generalize PSO to almost several solution representations in accepted and undemanding manner.

In GPSO, particle i's location is illustrated as vector $x_i = x_{i1}, x_{i2}, \ldots, x_{iN}$ taking each bit x_{ij} $(with\ j\ in\ \{1, N\})$ binary values 0 or 1. The main problem of GPSO is the particle movement formation.The proposed GPSO method is developed in this work operates as in below:

The pseudocode's first phase is the initialization process, in which the particles are carried out by means of the $S.Initialization()$ function, where S denotes Swarm. This special initialization method was adapted to gene selection as follows. The population was divided into four subsets of chromosomes such as particles initialized in different ways depending on the number of features in each particle. That is, less percentage of particles were initialized with prefixed value N selected genes located randomly. Another next to less percentage of particles were initialized with $2N$ genes, then 30% with 3N genes and finally, all other of particles (40%) were initialized randomly and 50% of the genes were turned on. In the GPSO experiments, N is equivalent to 4. In Algorithm shows the GPSO Pseudocode for Hamming space.

Algorithm: GPSO Pseudocode

Step 1: S ← Swarm Initialization()

Step 2: while not stop condition do

Step 3: for each particle x_i of the swarm S do

Step 4: evaluate(x_i)

Step 5: if f itness(x_i) is better than fitness(f_i) then

Step 6: $f_i \leftarrow x_i$

Step 7: end if

Step 8: if f itness(f_i)) is better than f itness(e_i) then

Step 9: $e_i \leftarrow f_i$

Step 10: end if

Step 11: end for

Step 12: for each particle x_i of the swarm S do

Step13:

$x_i \leftarrow$
$3Parent\ crossover\ Operator((x_i, w_1), (e_i, w_2), (f_i, w_3))$

Step 14: mutate(x_i)

Step 15: end for

Step 16: end while

Step 17: Output: best solution found

a second phase, after the evaluation of particles , historical and social position are updated. Finally, particles are "moved" by means of the 3 Parent crossover operator. And also, with a 10%probability a simple bit-mutation operator is applied for prevention of the early convergence. This will repeated until reach the final condition fixed to evolutions certain number.

4.2 Classification using KNN

K Nearest Neighbor (KNN) is a simple algorithm, which stores all cases and classifies new cases based on similarity measure. KNN algorithm also known as case based reasoning, instance based learning, memory based reasoning or lazy learning. KNN algorithms used in many applications like statistical estimation and pattern recognition etc., This is also non parametric classification method which has two types one is structure less NN techniques and another is structure based NN techniques. Nearest neighbor classification is used mainly when all the attributes are continues. Simple K nearest neighbor algorithm is shown in below

Steps 1) to discover the K training instances which are nearest to unknown instance

Step2) to select the most frequently occurring classification for these K instances.

Here this proposed leukemia gene classification method used efficient KNN technique for cancer and non-cancer cell classification from the given leukemia gene micro array dataset in most efficient manner which is discussed in detail as follows.

KNN based Leukemia gene Classification

Previously there are many algorithms designed for solving classification problems in machine learning have been applied to recent research classification of cancer especially luekemia with gene expression data. In this work a most efficient representative classification algorithm such as *k*-Nearest Neighbour, is applied to the classification. This is one of the most widespread methods among memory based orientation. Pearson's coefficient correlation and Euclidean , Manhattan and Minkowski distances have been used as the similarity measure. When this work have an input x and a reference set $M = \{m_1, m_2, \ldots, m_N\}$ the probability that x may belong to class c_j, $P(x, c_j)$ is defined as follows:

$$P(x, c_j) = \sum_{m_i \in KNN} Sim(x, m_i) P(m_i, c_j) - b_j \quad (10)$$

where $Sim(x, m_i)$ is the similarity between $x$and m_i and b_j is a bias term. Thus the KNN classifier is trained using the features selected, finally mostly determined features of the genes is accompanied to combine the cancer and non-cancer outputs of KNN classifier. After classification with some features is trained independently in classifier and produce their own outputs, finally get classification result will be judged by a combining module i.e. GPSO feature selection, where the most determined result of KNN classification is adopted.

5. Experimental Results and Comparison

Experimentation results of Leukemia Gene Classification is mainly relying on the dataset; to perform this proposed methodology take a leukemia 1 and leukemia 2 datasets with Acute Myelogenous Leukemia (AML), Acute Lympboblastic Leukemia (ALL) B-cell, and ALL T-cell, AML, ALL, and Mixed-Lineage Leukemia (MLL) diagnostic task type from Gene Expression Model Selector database [23] which is publicly available as open access. Both of these datasets provide information of micro array genes of both normal and cancer cells. The results of the proposed KNN classification with feature selection GPSO and without GPSO and existing classification algorithm is validated with use of leukemia cancer samples. The validation results of proposed KNN with GPSO and KNN with GPSO existing classification algorithm for GEMS leukemia 1 (Acute myelogenous leukemia (AML), acute Lympboblastic Leukemia (ALL) B-cell, and ALL T-cell) and leukemia 2 (AML, ALL, and mixed-lineage leukemia (MLL) datasets are used following metrics such as Sensitivity, Specificity, Accuracy and F-score in this work. The classification parameters definition and results of the proposed and existing classification results discussed is as follows.

Preprocessing Results

The missing data imputation results from the three normalization schemas is measured based on the root mean square error (RMSE) between the imputated gene feature values $\hat{x}_{ij}$ and the observered gene feature values x_{ij} from testing dataset ,RMSE is computed using the following equation ,

$$RMSE\left[\hat{x}_j\right] = \sqrt{\frac{\sum_{i=1}^{w_j}\left(\hat{x}_{ij} - x_{ij}\right)^2}{u_i}} \qquad (11)$$

u_i be the total number of the values or gene features imputated from normalization methods .

Preprocessing through Bi-PCA Results

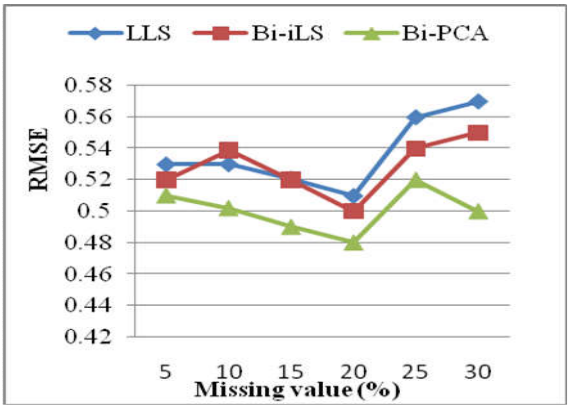

Figure 2: RMSE results for proposed and existing preprocessing method on leukemia dataset

Figure 2 shows the performance accuracy results of the preprocessing normalization schemas results .The proposed methods the normalization schema results is determined based on the RMSE between gene feature values and the observed gene feature values ,it shows that the proposed normalization methods produces less RMSE error values when compare to existing results .In the proposed work Bi-PCA produces best results in missing value rate imputation since it achieves less RMSE error value when compare to remaining methods for leukemia datasets.

Preprocessing through EICA

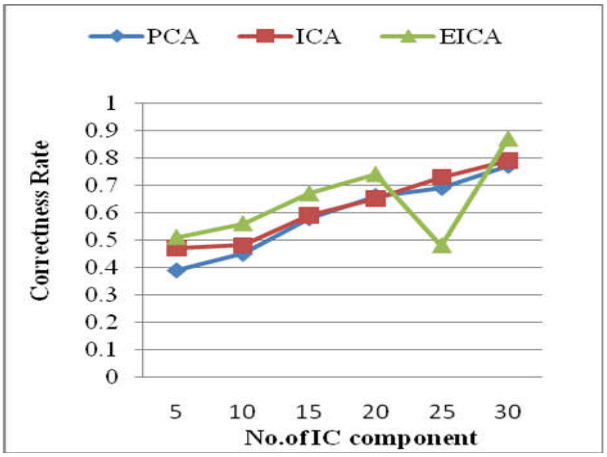

Figure 3. The results with applying proposed Pre-processing algorithm on microarray samples in Leukemia cancer data set

The aforementioned figure 3 clearly shows that the no. of independent IC components such as gene microarray data increases means, the correctness rate is also increased in frequent manner.

5.1 Classification through KNN with feature selection and without feature Selection

Sensitivity (S): Sensitivity is defined as the percentage of predicted and actual class which belongs to positive cases that were correctly identified, as determined using the equation:

$$\text{Sensitivity(S)} = \frac{T_1}{T_1+T_2} \quad (12)$$

=(Number of True Positive Assessment)/(Number of all Positive Assessment)

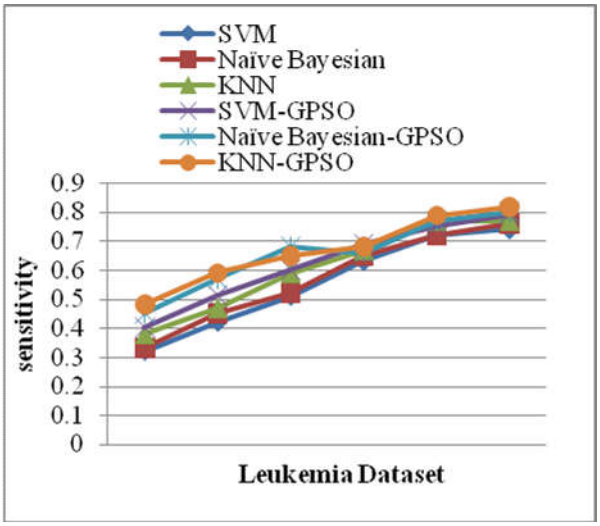

Figure 4: Sensitivity comparison vs. methods

The sensitivity results of proposed KNN and other existing classification methods with feature selection and without feature selection for leukemia dataset samples is illustrated in Figure 4. Sensitivity results of proposed KNN classification without GPSO is 0.77, in case of the addition of feature selection GPSO method into KNN classifier means the result also increased in is 0.049, this value is also higher than existing methods with or without GPSO. The feature selection GPSO is done to remove unimportant features in the preprocessed data. It is also applicable to all dataset samples where the result will be changed based on the characteristics of the dataset.

Specificity (Sp): Specificity is defined as the percentage of predicted and actual class which belongs to negative cases that were correctly identified, as determined using the equation,

$$\text{Specificity (Sp)} = \frac{F_2}{F_1+F_2} \quad (13)$$

=(Number of True Negative Assessment)/(Number of all Negative Assessment)

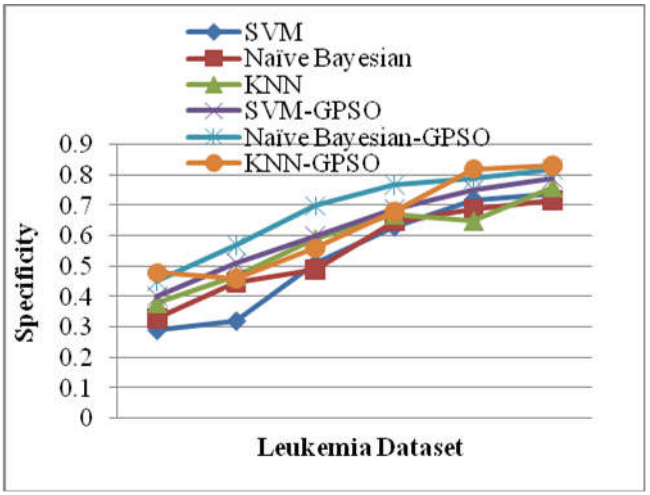

Figure 5: Specificity comparison vs. methods

Similarly specificity results of proposed KNN and existing classification methods are defined as the value of predicted and actual class which belongs to negative cases with GPSO and without GPSO, it shows that the proposed KNN methods have achieved higher value when it performs with GPSO, which is 0.072 higher than without GPSO and most of the other methods, which is shown in Figure.5, it is also shown that the proposed KNN and GPSO gives better performance than the other

methods specified in this work and this proposed methods perform well.

Classification Accuracy (A): Classification accuracy is defined as the percentage of the total amount of predictions and they are in both positive and negative cases that were correctly identified, as determined using the equation:

$$\text{Accuracy (A)}=\frac{T_1+F_2}{T_1+T_2+F_1+F_2} \quad (14)$$

=(Number of Correct Assessments)/Number of all Assessments)

Where, T_1 is the True Positive, F_2 is True Negative , and T_2 is False Positive and F_1 is False Negative

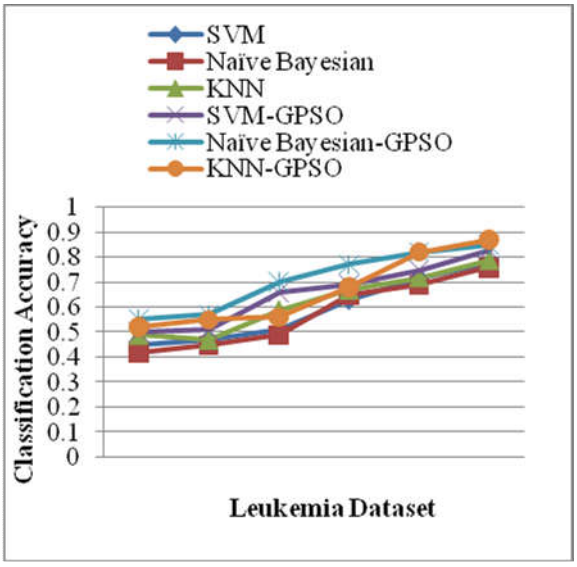

Figure 6: Accuracy comparison vs. methods

An accuracy of proposed KNN with GPSO classification is 0.04 value higher than the existing methods, the classification accuracy of the proposed and exiting method is increased when it performs with proposed GPSO feature selection, this will shows that it increases the classifier performance in the efficient manner, and also the proposed KNN also provide better result than the existing methods when it perform with GPSO the classification accuracy of the KNN gradually increased. So the test result shows that the contribution of the proposed work is more accurate, regardless positive is illustrated in Figure.6.

6. Conclusion and future work

In this proposed method shown the efficient result using KNN classifier with GPSO features selection technique, which was efficiently to classifies cancer and non-cancer cell using gene micro array leukemia dataset samples in most efficient way. The proposed technique consists of three stages, mainly, preprocessing, feature selection and classification using KNN classifier correspondingly. In the preprocessing stage, an EICA and Bi-PCA is proposed for analyze the given input dataset for further most accurate classification result. After that an important features of the leukemia gene data is selected by using GPSO, which are used in cancer classification process, then the cancer and non-cancer cells are classified based on the previous results using KNN classifier, finally the leukemia gene dataset is efficiently classified as cancer and non-cancer classes. According to experimental results show, the proposed gene classification method with use of the leukemia dataset, is efficiently classifies into malignant and benign classes. In future, it will be focused on an additional feature extraction process for most accurate classification results in proficient early cancer diagnosis.

References

1. A. Pease, D. Solas, E. Sullivan, M. Cronin, C. P. Holmes, and S. Fodor, "Light-generated oligonucleotide arrays for rapid dna sequence analysis," in *Proc. Natl. Acad. Sci.*, vol. 96, USA, 1994, pp. 5022–5026.
2. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, pp. 531–537, 1999.
3. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci*, vol. 96, pp. 6745–6750, 1999.
4. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L.

Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, pp. 530–536, 2002.

5. Juliusdottir, D. Corne, E. Keedwell, and A. Narayanan, “Two-phase EA/K-NN for feature selection and classification in cancer microarray datasets,” in *CIBCB*, 2005, pp. 1–8.
6. I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “Gene selection for cancer classification using support vector machines,” *Machine Learning*, vol. 46, no. 1-3, pp. 389–422, 2002. [Online]. Available: citeseer.ist.psu.edu/guyon02gene.html
7. J. Kohavi and G. H. John, “The wrapper approach,” in *Feature Selection for Knowledge Discovery and Data Mining*, 1998, pp. 33–50. [Online]. Available: citeseer.ist.psu.edu/article/kohavi97wrapper.html
8. Fix and J. L. Hodges, “Discriminatory analysis. nonparametric discrimination: Consistency properties,” 4, US Air Force School of Aviation Medicine, Randolph Field, TX, Tech. Rep., 1951.
9. C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.
10. Kai-Lin Tang, Wei-Jia Yao, Tong-Hua Li, Yi-Xue Li And Zhi-Wei Cao, ―Cancer Classification From The Gene Expression Profiles By Discriminant Kernel-Pls‖,Journal Of Bioinformatics And Computational Biology,Vol.8,Suppl.1(2010) 147-160.
11. Manuel Martin-Merino and Javier de las Rivas, "Kernel Alignment k-NN for Human Cancer Classification Using the Gene Expression Profiles", Springer link, Artificial Neural Networks – ICANN 2009.
12. Dimitris Maroulis, Dimitris Iakovidis, Ilias Flaounas and Stavros Karkanis, "A gene expression analysis system for medical diagnosis", IFIP International Federation for Information Processing, 2006.
13. Zhenyu Wang and Palade. V, "A Comprehensive Fuzzy-Based Framework for Cancer Microarray Data Gene Expression Analysis", Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, 2007.
14. Huang. D, Chow. T.W.S, Ma. E.W.M and Jinyan Li, " Efficient selection of discriminative genes from microarray gene expression data for cancer diagnosis", IEEE Transactions on Circuits and Systems, 2005.
15. url http://www.gems-system.org/.
16. Y. Cheng and G. M. Church, “Biclustering of expression data,” in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2000, pp. 93–103.
17. X. Gan, A. W. Liew, and H. Yan, “Discovering biclusters in gene expression data based on high-dimensional linear geometries,” *BMC Bioinform.*, vol. 9, no. 1, pp. 209–223, 2008.
18. W. C. Tjhi andL.Chen, “A partitioning based algorithm to fuzzy co-cluster documents and words,” *Pattern Recog. Lett.*, vol. 27, no. 3, pp. 151–159, 2006.
19. S. Das and S. M. Idicula, “Application of cardinality based grasp to the biclustering of gene expression data,” *Int. J. Comput. Appl.*, vol. 1, no. 18, pp. 47–54, 2010.
20. Z. Wang, C. W. Yu, R. C. C. Cheung, and H. Yan, “Hypergraph based geometric biclustering algorithm,” *Pattern Recog. Lett.*, vol. 33, no. 12, pp. 1656–1665, 2012.
21. Cheung YM, Xu L. An empirical method to select dominant independent components in ICA for time series analysis. Proceedings of the Joint Conference on Neural Networks. 1999:3883–7
22. A. Moraglio, C. D. Chio, and R. Poli, “Geometric Particle Swarm Optimization,” in 10th European conference on Genetic Programming (EuroGP 2007), ser. Lecture Notes in Computer Science, vol. 4445. Springer, Abril 2007.
23. A. Statnikov, I. Tsamardinos, Y. Dosbayev, C.F. Aliferis, "GEMS: A System for Automated Cancer Diagnosis and Biomarker Discovery from Microarray Gene Expression Data", International Journal of Medical Informatics, 2005 Aug;74(7-8):491-503.