MICROARRAY DATA CLASSIFICATION USING DUAL TREE M-BAND WAVELET FEATURES

Jayesh Manohar Sonawane,
School of Chemistry, Monash University, Clayton Campus,
Melbourne, Victoria, Australia 3800
Department of Energy Science and Engineering,
Indian Institute of Technology Bombay,
Mumbai, India-400076
Department of Chemical Engineering and Applied Chemistry,
University of Toronto, Toronto,
Canada, ON M5S
jayesh.sonawane@mail.utoronto.ca; jsonawane@iitb.ac.in

Shrihari D.Gaikwad,
Department of Instrumentation Engineering,
Smt. Indira Gandhi College of Engineering,
Navi Mumbai, India-40070
sdgaikwad.iitb@gmail.com

Gyan Prakash,
Software Engineer, R&D Department,
Vee Eee Technologies Solutions Pvt. Ltd,
Chennai, India
gyanprakash.95@gmail.com

Abstract: Deoxyribo Nucleic Acid (DNA) microarrays are widely used to monitor the expression levels of genes in parallel. It is possible to predict human cancer using the expression levels from a collection of DNA samples. Due to the vast number of genes expression level, it is challenging to analyze them manually. In this paper, data mining approach is used to extract the prevailing information from DNA microarray with the help of multiresolution analysis tool. Dual Tree M-Band Wavelet Transform (DTMBWT) is employed for the extraction of features from the given dataset at the 2nd level of decomposition. K-Nearest Neighbor (KNN) classifier is used for cancer classification. Results show that KNN classifier classifies five different cancer datasets; Breast, Colon, Ovarian, CNS, and Leukemia with over 90% accuracy.

Keywords: *DNA*, Microarray Data, DTMBWT, KNN.

I. INTRODUCTION

DNA microarray data analysis using Artificial Immune Recognition System (AIRS) is prepared in [1]. An improved version of the information gain feature selection method based on microarray data classification is performed in AIRS. A comparison is performed using three traditional classifiers that are K-NN, One-R, and Naive Bayes.

A breast cancer prediction system is discussed in [2] using clinical and DNA data of patients. Statistics based gene selection is employed at first. Support Vector machine (SVM) with different kernels and three Artificial Neural Network (ANN) algorithms are developed for cancer classification. SVM gives better classification than ANN for both clinical and DNA data.

ANN ensembles based gene expression data classification based on samples filtering is described in [3]. SVM, single ANN and bagging ANN is used to compare the outputs using leukaemia data sets. Ant Colony Optimization (ACO) is employed for gene selection in microarray dataset is discussed [4] for cancer classification. At first, the genes related to cancers are selected using ACO method. Then, the following classifiers such as SVM and multi-layer perceptron Neural Network (NN) are used for cancer classification.

Dimension reduction with Support Vector Regression (SVR) is explained in [5] for ovarian cancer DNA data. For a high dimensional problem like ovarian cancer DNA data, SVR with dimension reduction is employed to reduce the prediction errors, and it also avoids higher computational complexity. The dimension of ovarian DNA data is reduced from 9600 to 300 by SVR approach. It overcomes the block effect of DNA data. The number of reduced genes is based on the loss function of SVR.

The improvement of reliability of gene selection from genomic DNA data is presented in [6]. It employs 10-fold cross validation to access the repeatability and error of selected genes. To remove the effect of overwhelming data, only a small number of genes are selected using this method. Source variables are also identified from the selected genes.

A fuzzy expert system based on Genetic Swarm Algorithm (GSA) is discussed in [7] for microarray data classification. A near optimum rule set for the fuzzy expert system and tuning of membership function is achieved by GSA. An advanced and genetic operators related to a specific problem is employed to converge the GSA quickly.

Signal to Noise Ratio (SNR) and K-Means Clustering (KMC) based gene selection is discussed in [8]. For cancer classification, genetic programming is utilized. Informative features are selected using SND and KMC technique before classification.

DNA classification using Genetic Algorithm (GA) based classifier is explained in [9]. Usually, the DNA datasets consist of a small number of samples with a very large number of genes in each sample. The GA-based classifier is constructed using the selected features and its performance is analyzed using 10-fold cross validation.

Microarray data classification using cluster ensembles based on revisiting link is presented in [10]. Cluster ensembles are used followed by the construction of brief information matrix and transformed data. The link based cluster ensemble approach gives accurate clustering than state-of-the-art techniques.

Discrete Wavelet Transform (DWT) is employed in [11] for microarray data classification as a feature extraction technique with SVM classification. It includes approximation and detailed coefficients as features. It does not consider all detailed coefficients. Using maximum module method, useful details coefficients are selected.

DNA dataset is analyzed for cancer classification in [12]. It employs information gain based feature ranking for gene selection. Two and more than two class of cancer problems are analyzed. Among the various classification approaches such as decision tree, KNN, NN and Bayes, SVM classifiers outperform all approaches.

II. METHODS AND MATERIALS

The various expressions levels of genes are represented simultaneously by microarray data or DNA. In this study, the microarray data's are analyzed to classify them into normal or cancerous. Five microarray data sets of different cancers; Breast, Colon, Ovarian, CNS, and Leukemia are analyzed. DTMBWT is used to extract the features from the given microarray dataset, and KNN classifier is used to classify cancer. Figure 1 shows the microarray data classification system.

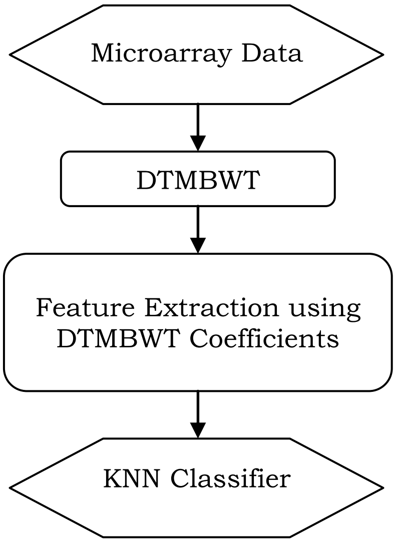

Fig. 1 Microarray data classification using DTMBWT and KNN

A. Dual Tree M-Band Wavelet Transform (DTMBWT)

Figure 2 shows the DTMBWT decomposition procedure which is a multiresolution analysis approach.

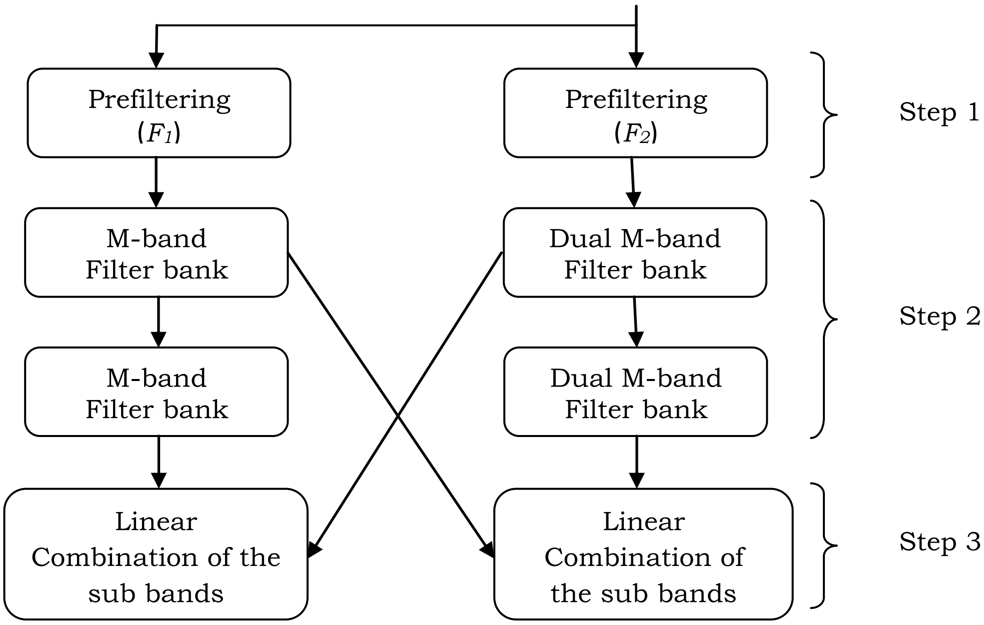

Fig. 2 M-band dual-tree decomposition system

Due to the multiresolution nature, the given signal is viewed at different resolution levels. More information can be available as features in these levels. These features are effectively analyzed for the classification. DTMBWT is proposed by Kingsbury in 2001 [13] and further analyzed by Selesnick in 2004 [14]. The process of decomposition consists of three steps; pre-filtering, M-band wavelet decomposition and direction extraction in the different sub bands.

B. K-Nearest Neighbor (KNN)

KNN classifier does not require any additional data for making the decision and the classification rules are computed based on the training samples only. Figure 3 shows an example of KNN classification. The decision is made based on the closest category of the training samples by the use of K value. It depends not only on the closest samples but also the value of K.

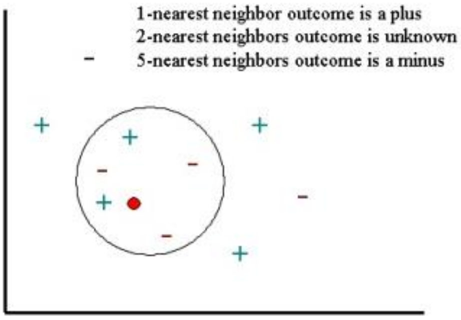
1-nearest neighbor outcome is a plus
2-nearest neighbors outcome is unknown
5-nearest neighbors outcome is a minus

Fig. 3 Example of K-NN classifier

III. RESULTS AND DISCUSSION

In this work, DTMBWT is used for feature extraction purpose, and classification is done by using KNN classifier. The application of wavelet transforms in the field of bioinformatics has been increasingly popular due to the following reasons; (i) it can represent the given data in multi-resolution and (ii) well-established localization in time-frequency space. In general, the performance is analyzed using k-fold validation approach. In this study, the number of fold is set to 10. Hence, the whole DNA microarray dataset is divided into ten subsets with an equal number of samples. Table 1 gives the description of database used in this study. The classification Accuracies vs. microarray datasets are graphically given in Figure 4.

TABLE 1 Microarray Database

Description	Types of cancer				
	Breast	Colon	Ovarian	CNS	Leukemia
#attributes	24481	1909	15154	7129	7129
#total cases	97	62	253	60	72
#normal cases	51	22	91	39	25
#abnormal cases	46	40	162	21	47

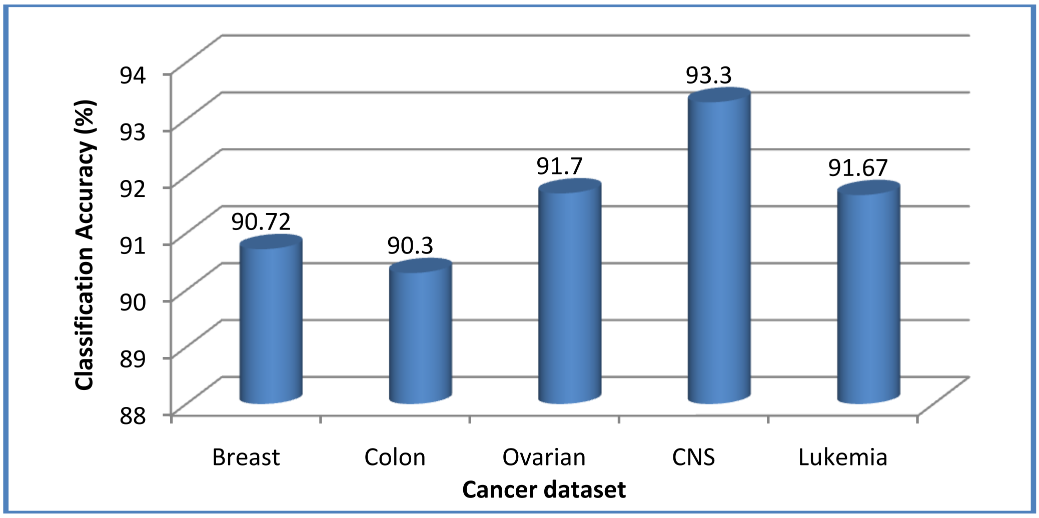

Fig. 4 Classification Accuracies vs. Microarray datasets

It is observed that KNN classifier provides more than 90% accuracies for all microarray datasets. Among the total cases, only 8 (Breast), 6 (Colon), 21 (Ovarian), 4 (CNS), and 6(Leukemia) cases are misclassified.

IV. CONCLUSION

In this paper, DTMBWT based microarray data classification is presented. Five different micro array datasets; Lung, cancer, prostate, colon and brain are chosen to evaluate the system. The dominant features are extracted using DTMBWT and are given to the classification stage using KNN classifier. As DTMBWT prove more selective in the frequency domain, all the DTMBWT coefficients are used as features. Experimental results on the standard benchmark DNA database of difference cancers show that the ensemble classification using SVM, NB, ANN, and DT outperforms the performance of the single classifier in terms of classification accuracy.

REFERENCES

[1]. C. Chen, C. Xu, R. Bie, and X.Z. Gao, “Artificial immune recognition system for DNA microarray data analysis”, IEEE Fourth International Conference on Natural Computation, Vol. 6, 2008, pp. 633-637.

[2]. A.H. Chen, G.T. Chen, J.C. Hsieh, and C.H. Lin, “BCPP: An intelligent prediction system of breast cancer prognosis using microarray and clinical data”, IEEE WRI World Congress on Computer Science and Information Engineering, Vol. 5, 2009, pp. 28-32.

[3]. W. Chen, H. Lu, M. Wang, and C. Fang, “Gene expression data classification using artificial neural network ensembles based on samples filtering”, IEEE International Conference on Artificial Intelligence and Computational Intelligence, Vol. 1, 2009, pp. 626-628.

[4]. Y.M. Chiang, H.M. Chiang, and S.Y. Lin, “The application of ant colony optimization for gene selection in microarray-based cancer classification”, IEEE International Conference on Machine Learning and Cybernetics, Vol. 7, 2008, pp. 4001-4006.

[5]. C.C. Chuang, S.F. Su, and J.T. Jeng, “Dimension reduction with support vector regression for ovarian cancer microarray data”, IEEE International Conference on Systems, Man and Cybernetics, Vol. 2, 2005, pp. 1048-1052.

[6]. L.M. Fu, and E.S. Youn, “Improving reliability of gene selection from microarray functional genomics data”, IEEE Transactions on Information Technology in Biomedicine, Vol.7, No.3, 2003, pp.191-196.

[7]. P.G. Kumar, T.A.A. Victoire, P. Renukadevi, and D. Devaraj, “Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm”, Expert Systems with Applications, Vol.39, No.2, pp.1811-1821.

[8]. S. Hengpraprohm, and P. Chongstitvatana, “Selecting Informative Genes from Microarray Data for Cancer Classification with Genetic Programming Classifier Using K-Means Clustering and SNR Ranking”, IEEE Frontiers in the Convergence of Bioscience and Information Technologies, 2007, pp. 211-218.

[9]. S. Hengpraprohm, S. Mukviboonchai, R. Thammasang, and P. Chongstitvatana, “A GA-Based classifier for microarray data classification”, IEEE International Conference on Intelligent Computing and Cognitive Informatics, 2010, pp. 199-202.

[10]. N. Iam-On, and T. Boongoen, “Revisiting link-based cluster ensembles for microarray data classification”, IEEE International Conference on Systems, Man, and Cybernetics, 2013, pp. 4543-4548.

[11]. S. Li, C. Liao, and J.T. Kwok, “Wavelet-based feature extraction for microarray data classification”, IEEE International Joint Conference on Neural Networks, 2006, pp. 5028-5033.

[12]. A. Osareh, and B. Shadgar, “Microarray data analysis for cancer classification”, IEEE 5th International Symposium on Health Informatics and Bioinformatics, 2010, pp. 125-132.

[13]. N. Kingsbury, “Complex wavelets for shift invariant analysis and filtering of Signals”, Applied and computational harmonic analysis, Vol. 10, No. 3, 2001, pp. 234-253.

[14]. I.W. Selesnick, “The double-density dual-tree DWT”, IEEE Transaction on Signal Processing, Vol. 52, No. 5, 2004, pp. 1304-1314.