International Journal of Scientific and Technical Advancements
ISSN: 2454-1532

A Survey related to Gene Selection and Cancer Classification using Relevance Vector Machine (RVM)

S. Sasikala
Department of Computer Science, Sree Saraswathi Thyagaraja College, Pollachi, Tamilnadu, India-642 002
Email address: sasivenkatesh04@gmail.com

Abstract— Now a day's cancer is the most dangerous diseases in the world. There were bunch of proposal from a variety of establishers and detailed picture examination was still under processing. Generally cancer is defined as the abnormal growth and uncontrolled growth in the human bodies. Many cells are constructed to form a living organisms like planets, humans and animals etc. Each cell containing one organism. The human body having constructed by cells most of the cells having less periodic life time and it can changed in a cyclic manner. Some of the abnormal time the malignant cells will growth in a certain period and it will produce the cancer cell in human body. The malignant is the beginning stage of cancer. On the odd occasion, there is a collapse of the mechanism and a potentially malignant cell survives, replicates and cancer is the consequence. Some amount of cancer is obviously rooted. Heredity habit donates to a quantity segment of fully grown cancers. The huge number of cancers is caused mainly due to the eating habit, their working situations, viruses, bacteria, artificial radiation and chemicals. These are generally known as "environmental" risk factors for cancer.

Keywords— Cancer classification; gene selection; neural network; SVM; machine learning techniques; classifier; RVM; gene expression.

I. INTRODUCTION

Machine learning is a regimen discipline that works with the design and implementation of algorithms that facilitates machines to build up behaviors based on empirical data. A learner can take benefits of the past data to obtain features of interest of their unknown basic probability distribution. Data is seen as examples that show relations between observed variables. The main aim on machine learning research is to make them learn automatically to identify complex patterns and make clever decisions based on the data. But the complexity lies in the fact that the possible inputs are too huge to be covered by a group of training data. The paper is organized as follows. Section II discusses the literature review related to RVM. Section III discusses the scope for future work. Section IV concludes the paper with discussions.

II. RELATED WORK

A. Cancer Classification using Neural Networks

Ahmad M. Sarhan suggested the cancer classification based on micro array gene expression data using DCT and ANN. The author mainly deals about, a stomach cancer detection system based on Artificial Neural Network (ANN), and the Discrete Cosine Transform (DCT), is developed. The present model uses DCT to obtain classification features from stomach microarrays. The obtained features from the DCT coefficients are then employed to an ANN for classification (tumor or non tumor). The microarray images were collected from the Stanford Medical Database (SMD). Simulation results showed that the proposed model achieves a higher success rate. DNA Microarrays are glass microscope slides onto which genes are attached at fixed and ordered locations. Each gene series is recognized by a location of a spot in the array. The DNA is spotted directly using a microarray printer onto the slide. It is possible to study a gene expression within a single sample or to evaluate gene expressions inside two tissue models, such as in tumor and non tumor tissues. Thus, a strong model for stomach cancer detection using microarrays is presented by the author. The system involves a feature extraction stage and an ANN classification stage. The feature extraction stage employs the 2D DCT to condense the input microarray. Low frequency parts of the DCT array comprises of most of the energy/information of the input microarray. These parts were, used as typical features and were obtained using a windowing approach. The author also examines through simulations, optimal parameters such as the optimal number of DCT coefficients/features and the best ANN structure for the detection of stomach cancer. The proposed method produces a success rate of 99.7%. The sensitivity, specificity, and accuracy of the system were found to be equal to 99.2%, 100%, and 99.66% respectively. Experimental tests on the SMD Database achieved 99.7% of recognition accuracy using only100 DCT coefficients, with a simple 2-layer ANN structure and low computational cost.

Wooten et al., in recommended a numerical examination for the enhancement ANN in detection of breast cancer using a planer broadband antenna and a three-region breast technique. In this proposed technique, a Modified Four point antennas are used for constructing several wave polarizations. The result of wave polarization on statistical detection is fully described in this approach by the author.

Mass spectrometry-based proteomics provides a significant approach for the efficient diagnosis of different diseases. However, there are some issues in the mass spectral data such as huge volume, data complexity and the presence of noise which make the investigation of the proteomic pattern very tough. A neural by Xu et al., in presented a neural network

S. Sasikala, "A survey related to gene selection and cancer classification using relevance vector machine (RVM)," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 203-207, 2016.

based approach for proteomic pattern analysis for prostate cancer screening. The proposed approach consists of three stages namely feature selection depending on statistical significant test, classification by a Radial Basis Function Neural Network (RBFNN) a Probabilistic Neural Network (PNN), and ultimately results in optimization via ROC analysis. The experimental observation shows that the proposed approach is very effective when compared with the existing approaches. The proposed approach has high sensitivity (97.1%) and specificity (96.8%) when combined with prostatic biopsy and is expected to help in early detection of prostate cancer.

Fooladi proposed that the sensitivity to the induction of chromosomal damage by ionizing Gama Exposure is higher in breast cancer patients than in normal healthy controls. The gamma effect in each person's lymphocytes and the comparison between two groups is examined, seventy two hours after blood sample culturing, by exposing the samples to gamma rays and then they are harvested. The exposure of gamma rays cause abnormality in chromosomes. The database used in this proposed approach includes chromosome breakage in seven chromosome groups and age of patients. In this technique Principle Component Analysis (PCA) is used for feature selection stage. Then Artificial Neural Networks (ANN) is used for classification of normal cases from abnormal cases. The experimental observation shows that the result is obtained with an accuracy rate of 93.09% for Neural Networks (NN) classifier.

A novel texture analysis approach based on fuzzy co-occurrence matrix concept is proposed by Cheng et al., in. This approach is used to handle early and exact breast cancer diagnosis by examining the microscope-slide biopsy images. A novel feature extraction algorithm is used to extract the features from the digitized images, and then the extracted features are given as input to a multilayer back-propagation neural network to categorize the images into three risk groups. The performances of the conventional cancer diagnosis methods and the proposed algorithm are evaluated and it is found that, this approach has higher performance compared to the existing methods. The proposed technique has wide applications in the areas of pattern recognition and image processing.

Shukla, in recommended a novel technique to simulate a Knowledge Based System for diagnosis of Breast Cancer using Soft Computing tools like Artificial Neural Networks (ANNs) and Neuro Fuzzy Systems. The feed-forward neural network has been trained using three ANN algorithms namely

(1) Back Propagation Algorithm (BPA)
(2) Radial Basis Function (RBF) Networks and the Learning Vector Quantization (LVQ) Networks
(3) Adaptive Neuro Fuzzy Inference System (ANFIS)

The simulation is done using MATLAB and performance is evaluated by considering the metrics like accuracy of diagnosis, training time, number of neurons, number of epochs etc. The simulation results show that this proposed approach can be very effective for early detection of Breast Cancer which also helps oncologists to enhance the survival rates significantly.

A vital early sign of breast cancer are Clusters of micro calcifications in mammograms. Song yang Yu Ling Guan in described a Computer-Aided Diagnosis (CAD) system for the automatic detection of clustered micro calcifications in digitized mammograms. The approach consists of two main steps.

(1) In order to segment the potential micro calcification pixels in the mammograms mixed features consisting of wavelet features and gray level statistical features are used, and labeled into potential individual micro calcification objects by their spatial connectivity.
(2) A set of 31 features extracted from the potential individual micro calcification objects are used for the detection of the Individual microcalcifications. The biased power of these features is analyzed using general regression neural networks via sequential forward and sequential backward selection techniques.

Multilayer feed forward neural networks are the classifiers used in these two steps. Nijmegen database of 40 mammograms containing 105 clusters of microcalcifications is used in this experimentation. The performance is evaluated using the Free-Response Receiver Operating Characteristics (FROC) curve. The experimental observation shows that, the proposed approach gives significant detection performance. 90% mean true positive detection rate is obtained at the cost of 0.5 false positive per image when mixed features are used in the first step and 15 features selected by the sequential backward selection method are used in the second step.

One of the main computer-aided mammographic breast cancer detection techniques is the Mass detection. In order to reduce mortality, early detection of primary tumor is very vital. Computer-Aided Diagnosis approach is very much useful for radiologist in detecting and diagnosing abnormalities earlier and faster than conventional screening techniques. Jasmine et al., in proposed a new approach for detecting micro calcification in digital mammograms by using the combination of wavelet analysis of the image by applying Artificial Neural Networks (ANN) for building the classifiers. The micro calcification belongs to high frequency components and the detection of micro calcification is obtained by extracting the micro calcification features from the wavelet analysis of the image and these results are used as an input of neural network for classification. The neural network consists of one input, two hidden and one output. The experimental observation shows that the proposed approach can provide true detection rate approximately 87% and zero false detection per image which is significant. The proposed approach evaluation is carried on Mammography Image Analysis Society (MIAS) dataset.

B. Cancer Classification using Machine Learning Techniques

Several machine learning techniques are used for the classification of cancer. This section discusses about the cancer classification techniques which uses the machine learning approaches like Support Vector Machine (SVM),

S. Sasikala, "A survey related to gene selection and cancer classification using relevance vector machine (RVM)," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 203-207, 2016.

Extreme Learning Machine (ELM) and Relevance Vector Machine (RVM).

C. Cancer Classification using SVM

S. Ramaswamy [1] described about multiclass cancer diagnosis through tumor gene expression signatures, which purposely says about, the complex combination of clinical and histopathological data for optimal treatment of patients with cancer depends on establishing accurate diagnoses; it seems to be difficult because of atypical clinical presentation or histopathology. To determine whether the identification of multiple common adult malignancies could be achieved purely by molecular classification, for example the author, subjected 218 tumor samples, spanning 14 common tumor types, and 90 normal tissue samples to oligonucleotide microarray gene expression analysis. Here by using SVM the accuracy of multi class is predicted by expressing 16,063 genes and sequence tags. So this had an output of 78%, much greater than the accuracy of random classification that is about 9%. In recent times, DNA microarray-based tumor gene expression profiles have been used for cancer diagnosis. Anyhow, studies have been limited to few cancer types and have spanned multiple technology platforms complicating comparison among different datasets. The possibility of cancer diagnosis across all of the common malignancies based on a single reference database has not been explored. For a sample 314 tumors and 98 normal tissues were considered, in that 218 tumor and 90 normal tissue samples passed quality control criteria and were used for subsequent data analysis. The remaining 104 samples of the data either failed quality control measures of the amount and quality of RNA, as assessed by spectrophotometric measurement of OD and agarose gel electrophoresis, or yielded poor-quality scans. Scans were discarded if mean chip intensity exceeded 2 SDs from the average mean intensity for the whole scan set, if the proportion of ‘‘present’’ calls was less than 10%, or if microarray artifacts were visible. The problem of biological and measurement noise, contaminating nonmalignant tumor components, and inclusion of genetically heterogeneous samples within clinically defined tumor classes may all effectively decrease predictive power in the multiclass setting. Increased gene number likely allows for accurate prediction despite these factors. A greater variety and large number of tumors with detailed clinic pathological characterization will be required to fully explore the true limitations of gene expression-based multiclass classification.

L. wang et al., [2] proposed the accurate cancer classification using expression of very few genes, the author aim at finding the smallest set of genes that can ensure highly accurate classification of cancers from microarray data by using supervised machine learning algorithms. The importance of finding the minimum gene subsets is three-fold: 1) It greatly reduces the computational burden and “noise” arising from irrelevant genes. From the examples stated in this approach, finding the minimum gene subsets even allows for extraction of simple diagnostic rules which lead to accurate diagnosis without the need for any classifiers. 2) The gene expression tests are simplified to include only a very small number of genes rather than thousands of genes, which can bring down the cost for cancer testing significantly. 3) It calls for additional investigation into the possible biological relationship between these small numbers of genes and cancer development and treatment. The simple yet very effective method involves two steps. In the first step, the author chooses some important genes using a feature importance ranking scheme. In the second step, the author tests the classification capability of all simple combinations of those important genes by using a good classifier. For three “simple” and “small” data sets with two, three, and four cancer (sub) types, the approach obtained very high accuracy with only two or three genes. For a “large” and “complex” data set with 14 cancer types, the author divided the whole problem into a group of binary classification problems and applied the 2-step approach to each of these binary classification problems. Through this “divide-and-conquer” approach, the author obtained accuracy comparable to previously reported results but with only 28 genes rather than 16,063 genes. In general, this method can significantly reduce the number of genes required for highly reliable diagnosis by the technique of SVM-T test analysis. The author analyzed finally and gave the accuracy rate of 100% by three combinational iteration techniques.

D. Cancer Classification using RVM

Liyang et al., [11] presented Relevance Vector Machine (RVM) technique for reorganization of MCs in digital mammograms. RVM [12] depends on the Bayesian assessment theory, of which a unique feature is that it can provide a thin decision function defined by only a least numbers called relevance vectors. By using this sparse property of the RVM, the author exploits computerized reorganization approaches that are both accurate and computationally efficient for MC detection in mammograms. The author suggests MC detection as a supervised-learning issue and applied RVM classifier to find out the presence at each location in the mammogram if an MC objects. With the purpose of increasing the calculation speed additional, the author developed a two-stage classification network, in which a computationally much easier linear RVM classifier is applied initially to rapidly remove the overwhelming majority, non-MC pixels in a mammogram from any advance consideration. This approach by Liyang is estimated using a database of 141 clinical mammograms (all containing MCs), and evaluated against a well-tested support vector machine (SVM) classifier [16]. The performance of the detection is tested with the application of Free-response Receiver Operating Characteristic (FROC) curves. It is illustrated in the test that the RVM classifier could considerably reduce the computational complexity of the SVM with better detection accuracy. In particular, the two-stage RVM approach reduces the detection time from 250 s for SVM to 7.26 s for a mammogram (approximately 35-fold reduction). As a result, the RVM classifier by Liyang found to be more beneficial for real-time processing of MC clusters in mammograms.

W. Zhang et al., [14] puts forth a novel approach for the multicategory cancer classification. SVM-RFE is a vital technique of the gene selection approaches, which integrates SVM with recursive feature removal, and the technique ranks

S. Sasikala, “A survey related to gene selection and cancer classification using relevance vector machine (RVM),” *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 203-207, 2016.

the genes with recursive process. A novel machine technique called RVM is proposed by Tipping [15] in 2000, as an option to SVM. The authors described RVM-RFE approach for gene selection by integrating RVM and RFE. The performance evaluation on the real datasets reveals that RVM-RFE can lead to significant accuracy and shorter running time. Thus, this approach is also much better than linear RVM and other conventional approaches.

Carin et al., uses Relevance Vector Machine [17] feature selection and classification for underwater targets. Song et al., [18] gives a greedy algorithm for gene selection [KIP05, 19] based on SVM and correlation. Microarrays serve as a potential and competent tool to study thousands of genes and examine their activeness in normal or cancerous tissues. Usually, microarrays are used to evaluate the expression levels of thousands of genes in a cell mixture. Gene expression data obtained from microarrays can be used for several applications. Gene selection is equivalent to the feature selection problem addressed in the machine-learning field. In a nutshell, gene selection is the difficulty of recognizing a least collection of genes that are accountable for some events (for instance the occurrence of cancer). Informative gene selection is an essential problem taking place in the investigation of microarray data. A novel algorithm is presented for gene assortment that integrates Support Vector Machines (SVMs) [20] with gene correlations. Experimental observation reveals that the new algorithm, known as GCI-SVM, acquires a higher classification accuracy using a lesser number of chosen genes than the recognized algorithms in the literature.

Establishing transcriptional regulatory networks by investigation of gene expression data and promoter sequences confirms great promise. Y. Li et al. [21] developed a novel promoter classification technique by means of a Relevance Vector Machine (RVM) and Bayesian statistical principles to recognize discriminatory characteristics in the promoter sequences of genes that can accurately classify transcriptional responses. The technique was implemented to microarray data acquired from Arabidopsis seedlings treated with glucose or abscisic acid (ABA). Genes which has >2.5-fold changes in expression level, ~70% were accurately predicted as being up- or down-regulated (based on 10-fold cross-validation), based on the presence or absence of a small set of discriminative promoter motifs. Majority of these motifs have fixed regulatory functions in sugar- and ABA-mediated gene expression. One promoter motif that was not recognized to be took part in glucose-responsive gene expression was recognized as the powerful classifier of glucose-up-regulated gene expression. Confers glucose-responsive gene expression is combined with an additional promoter motif, thus confirming the classification technique. This work establishes a comprehensive model of glucose and ABA transcriptional regulatory networks and their communications, which will assist in recognizing the methods linking metabolism with development in Arabidopsis. This investigation demonstrates that machine learning approaches coupled to Bayesian statistical techniques hold considerable promise for recognizing functionally considerable promoter sequences.

In many real-world classification difficulties the input comprises of a huge number of potentially inappropriate characteristics. Y. (Alan) Qi et al., [22] proposed a new Bayesian framework for finding out the importance of input features. This technique broadens one of the most successful Bayesian methods for feature selection and sparse learning, recognized as Automatic Relevance Determination (ARD). ARD discovers the relevance of features by optimizing the model marginal likelihood, also called as the evidence. This leads to over fitting. To ignore this difficulty, predictive ARD was presented depending on estimate the predictive performance of the classifier. On the other hand, the actual leave-one-out predictive performance is commonly very expensive to compute, the expectation propagation (EP) algorithm developed by Minka [22] provides an approximation of this predictive performance as a consequence of its iterations. It is exploited in this algorithm to do feature collection, and to select data points in a sparse Bayesian kernel classifier. In addition, two other enhancements were made to previous approaches, by substituting Laplace's estimation with the commonly more correct EP, and by integrating the fast optimization algorithm proposed by Faul and Tipping [23]. The experiments observations revealed that this approach is based on the EP approximation of predictive performance are very precise on test data than relevance determination by optimizing the evidence.

III. Scope For Future Research

Among several algorithms, Machine learning algorithms were found to be more effective for Gene Selection and Cancer Classifications. Analytical network processing (ANP) can be applied for effective Gene selection. Machine learning based methodologies can be used to improve accuracy for the process of cancer classifications.

IV. Conclusion

In the field of Bioinfomatics, Cancer classification is found to more vibrant research area. In this survey we discussed about various machine learning techniques for gene selection and cancer classification. The important machine learning algorithms like Support vector machine, Extreme learning machine and Relevance vector machine were used. Among all, the RVM approach is found to be more effective in terms of accuracy in cancer diagnosis.

References

[1] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. H. Yeang, M.Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M. Loda, E. S. Lander, and T. R. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceeding Nat'l Academy Sciences*, USA, vol. 98, no. 26, pp. 15149-15154, 2002.

[2] L. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol.4, no. 1, pp.

[3] R. Zhang, G. B. Huang, N. Sundararajan, and P. Saratchandran, "Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis," vol. 4, no. 3, 2007.

S. Sasikala, "A survey related to gene selection and cancer classification using relevance vector machine (RVM)," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 203-207, 2016.

[4] G. B. Huang and C. K. Siew, "Extreme learning machine: RBF network case," *International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pp. 1029-1036, 2004
[5] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for classification of tumors using gene expression data," *J. Am. Statistical Assoc*., vol. 97, no. 457, pp. 77-87, 2002.
[6] D. Serre, "Matrices: Theory and applications," *Springer-Verlag*, 2002.
[7] G. B. Huang, L. Chen and C. K. Siew, "Universal approximation using incremental constructive feed forward networks with random hidden nodes," *IEEE Transactions on Neural Networks*, vol. 17, no. 4, pp. 879–892, 2006.
[8] M. Ringner, C. Peterson, and J. Khan, "Analyzing array data using supervised methods," *Pharmacogenomics*, vol. 3, no. 3, pp. 403-415, 2002.
[9] T. Helmy and Z. Rasheed, "Multi-category bioinformatics dataset classification using extreme learning machine," *IEEE Congress on Evolutionary Computation (CEC '09)*, pp. 3234–3240, 2009.
[10] J. Sanchez-Monedero, M. Cruz-Ramirez, F. Fernandez-Navarro, J. C. Fernandez, P. A. Gutierrez, and C. Hervas-Martinez, "On the suitability of extreme learning machine for gene classification using feature selection," *Intelligent Systems Design and Applications - ISDA*, pp. 507-512, 2010.
[11] L. Wei, Y. Yang, R. M. Nishikawa, M. N. Wernick, and A. Edwards, "Relevance vector machine for automatic detection of clustered microcalcifications," *Medical Imaging, IEEE Transactions*, vol. 24, no. 10, pp. 1278–1285, 2005.
[12] S. Chen, S. R. Gunn, and C. J. Harris, "The relevance vector machine technique for channel equalization application," *IEEE Trans on Neural Networks*, vol. 12, no. 6, pp. 1529-1532, 2001.
[13] Y. Lee and C. K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, pp. 1132-1139, 2003.
[14] W. Zhang and J. Liu, "Gene selection for cancer classification using relevance vector machine," *Bioinformatics and Biomedical Engineering, ICBBE*, pp. 184–187, 2007.
[15] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research I*, pp. 211-214, 2001.
[16] Y. Lee and C. K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, pp. 1132-1139, 2003.
[17] M. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research I*, pp. 211-214, 2001.
[18] M. Song and S. Rajasekaran, "A greedy algorithm for gene selection based on SVM and correlation," *International Journal of Bioinformatics Research and Applications - IJBRA*, vol. 6, no. 3, pp. 296-307, 2010.
[19] M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, and M. A. Koval, "Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning," *Nature Medicine*, vol. 8, no. 1, pp. 68-74, 2002.
[20] El-Naqa. I, Y. Yang, M. N Wernick, N. P Galatsanos, and R. Nishikawa, "Support vector machine learning for detection of microcalcifications in mammograms," *IEEE International Symposium on Biomedical Imaging*, pp. 201–204, 2002.
[21] Y. Li, K. K. Lee, S. Walsh, C. Smith, S. Hadingham, K. Sorefan, G. Cawley, and M. W. Bevan, "Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a relevance vector machine," *Genome Research*, vol. 16, no. 3, pp. 414-427, 2006.
[22] Y. (Alan) Qi, T. P. Minka, R. W. Picard, and Z. Ghahramani, "Predictive automatic relevance determination by expectation propagation," *Proceeding ICML '04 Proceedings of the twenty-first international conference on Machine learning*, 2004.
[23] A. C. Faul and M. E. Tipping, "Analysis of sparse bayesian learning," *Neural Information Processing Systems-NIPS*, pp. 383-389, 2001.

S. Sasikala, "A survey related to gene selection and cancer classification using relevance vector machine (RVM)," *International Journal of Scientific and Technical Advancements*, Volume 2, Issue 1, pp. 203-207, 2016.