Volume 6, Issue 4, April 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Support Classifier Based Genetic Algorithm for Cancer Cell Detection using Feature Reduction

Taniya Parmar
M.Tech Scholer CSE Department, VNS College,
Bhopal, India

Amit Kumar Nandanwar
A.P. CSE Department, VNS College,
Bhopal, India

Abstract— Cancer is one among the key analysis topics within the medical field. Associate correct detection of various cancer tumor varieties has nice worth in providing higher treatment facilities and risk minimisation for patients. Recently, dna microarray-based organic phenomenon profiles are utilized to correlate the clinical behavior of cancers with the differential gene expression levels in cancerous and benign tumors. Associate correct classifier with linguistic interpretability employing a small range of relevant genes is useful to microarray knowledge analysis and development low rate diagnostic tests. many well-known and often used techniques for coming up with classifiers from microarray knowledge, like a support vector machine, neural networks, k-nearest neighbor, and logistical regression model, suffer from low quality. This work reviews several basic algorithms found among the literature and assesses their performance during a} very controlled state of situation. to resolve this type of issues genetic rule approach uses for feature choice and parameters optimisation. This work involves detailed study and coming up with a framework that includes genetic rule with SVM for feature choice and classification on the training dataset. identification of diseases is incredibly vital and tough task, and this study helps for locating technique that performs feature choice and parameters setting. the target of this work is to analysis existing work for feature set and mentioned technique for rising classification accuracy.

Keywords— Cancer, dna microarray, feature choice, gene expression

.

I. INTRODUCTION

Microarray may be a valuable technique for measurement expression knowledge of thousands of genes at the same time. a very important rising medical application domain for microarray gene expression identification technology is medical call support within the type of identification of sickness additionally as prediction of clinical outcomes in response to treatment. the two areas in medication that presently attract the foremost attention in this respect are management of cancer and infectious diseases [1].The prediction of the diagnostic class of a tissue sample from its expression array constitution exploitation the microarray knowledge from tissues in known classes well known as classification. The samples are sometimes the experiments and therefore the classes are the kinds of tissue samples. variety of systematic strategies are developed and studied to classify cancer varieties exploitation gene expression knowledge [2].

Computational analysis and computing will facilitate researchers to collate a group of signature genes for a particular illness [3, 4]. Since the value of microarray chips is incredibly high and additionally we've got not enough tissue samples from cancer patients, the quantity of records in microarray datasets is typically too few, that isn't appropriate for many machine learning algorithms. additionally, the process and material used for microarray analysis is often completely different between makers. Therefore, it's tough to identify a novel set of genes which will kind associate integrated dataset [5]. Moreover, whereas the quantity of samples for every cancer sort is typically balanced for electronic device analysis, the ratio of cancer patients to traditional adults is usually a lot of smaller within the real world.

Since the quantity of genes is usually a lot of larger than the quantity of tissue samples [6], one among the most challenges would be selecting little and discriminative set of effective genes among tens of thousands of genes, that may be a terribly troublesome task. Therefore, gene choice becomes the foremost necessary necessity for a microarray based mostly cancer system. However, the most effective combination of classification and gene choice is known poorly, as a result of there's another method bother related to training microarray knowledge. this is often the matter of ‘‘over-fitting’’[7]. In short, over-fitting means one will get sensible performance employing a training set, however once new data set is used, a satisfactory result can't be obtained exploitation the trained model. this happens usually after we have a small variety of high-dimensional samples (for each training and testing the model). sadly, we've got precisely such a problem in cancer tumor detection exploitation micro-array datasets.

II. SOFT COMPUTING METHODS

Classification & clustering is a method in which Objects are characterized by one or more features Classification is a task which assigns objects to classes or groups on the basis of measurements made on the objects

– Have labels for some points

– Want a “rule” that will accurately assign labels to new points
– Supervised learning Clustering is to group observations that are “similar” based on predefined criteria.
– No labels
– Group points into clusters based on how “near” theyare to one another
– Identify structure in data
– Unsupervised learning

Table I Soft computing clustering and classification [8].

Various Important Clustering Methods	**Various important Classifiers**
Hierarchical Methods	Supervised Methods
• Agglomerative hierarchical clustering • Divisive hierarchical clustering • Single-link clustering • Complete-link clustering • Average-link clustering	• Naïve Bayes Classifier • J48 Decision Trees • Support Vector Machines
Partitioning Methods	Unsupervised method
• Error Minimization Algorithms. • Graph-Theoretic Clustering	• SenseClusters (an adaptation of the K-means clustering algorithm)
Density Model-based Clustering Methods	Instance-based learning
• Decision Trees. • Neural Networks	• Nearest neighbor classifier
Grid-based Soft-computing Methods	Perceptron-based techniques
• Fuzzy Clustering • Evolutionary Approaches for Clustering • Simulated Annealing for Clustering	• Single layered perceptrons • Multilayered perceptrons • Radial Basis Function (RBF) networks
	Statistical learning algorithms
	• Naive Bayes classifiers • Bayesian Networks • Instance-based learning

III. RECENT FEATURE SELECTION TECHNIQUES

A. Selected Choices Exploitation Fuzzy Modeling

The paper [9] has conferred a study of medical processing and data processing that's involving the employment of 11 feature choice ways and three fuzzy modeling ways; such strategies are not all offered during a business processing and data processing package. the target is to see that that combination of feature choice and fuzzy modeling ways has the most effective performance for a given dataset. two medical datasets and one industrial dataset were tested with multiple stratified cross-validation. All the mix of feature choice and fuzzy modeling ways were applied.

B. Feature Choice Methodology Supported Association Rules

The paper [10] a hybrid methodology for identification of erythemato-squamous diseases supported Association Rules (AR) and Neural Network (NN). Feature extraction is t the key for pattern recognition and classification. If the options are not chosen well, the most effective classifier will perform poorly. A feature extractor ought to scale back the feature vector to a lower dimension that contains most of the helpful knowledge from the initial vector. So, AR is utilized for reducing the dimension of erythemato-squamous diseases dataset and NN is utilized for intelligent classification. The projected AR+NN system performance is compared with NN model. The dimension of input feature space is reduced from thirty three to twenty four by exploitation AR. In testing stage, proposed system performances square measure evaluated by applying 3-fold cross validation methodology to the erythemato-squamous diseases dataset. The classification rate of planned system is 98.61% for twenty-four inputs. This analysis demonstrate that the AR are going to be used for reducing the dimension of feature vector and planned AR+NN model are going to be used for getting better automatic diagnostic systems for different diseases.

C. Gene Choice Multi-View Fitness Function

The paper [11] provided a CD-MFS formula that relies on memetic biological process concept that uses correct set of fuzzy if-then rules that may classify gene expression knowledge. It begins with less quality rules, and ends up in top quality rule set. This formula classifies cancerous and benign tumors with efficiency and has acceptable accuracy. 14_Tumors cancer dataset was evaluated by our projected formula and it compared with different classification systems. Results indicate that our CD-MFS outperforms several well-known and up so far classifications. Moreover, the paper suggests new reasoning technique and Multi-View fitness functions in memetic algorithms. The introduced Multi-View fitness functions classifying cancerous tumors from organic phenomenon knowledge by considering each native and world fuzzy rule strength. This work tend to in addition targeted on manufacturing significant fuzzy rules from the memetic recursive, that square measure additional explainable for a doctor. On the opposite hand, every reasonably neoplasm is clearly distinguishable by if-then rules made by the formula.

D. Feature Generation Mistreatment Genetic Programming

This paper [12] presents a genetic programming primarily based methodology to classify diabetes knowledge. To facilitate the choice of options and for evaluating the effectiveness of diabetes options numerous methodologies are utilized in this analysis. By making combos of selected features gp has been accustomed modify the method of generating new options. A variation of gp that is named gp with cps and it's been used that performs higher than the quality gp.GP improves the performance and it reduces the eight dimensions to single dimension. The new options that generated by physician square measure tested by KNN and SVM to judge the performance and also the results demonstrate that physician generated options show important improvement in performance as compared to the performance achieved by original diabetes options. as compared with different ways gift within the literature shows the prevalence of the given technique.

E. Feature Selection Technique Based Mostly Hybrid Intelligent System

Ovarian cancer diagnosis could be a very important study as a result of early detection and accuracy staging square measure the keys to extend the survival rate of the patient. In papers, [13] propose a completely unique hybrid intelligent system, that derives easy however convincing fuzzy inference rules to diagnose ovarian cancer and confirm its stage in keeping with the amount of seriousness. Our given self-organizing model is understood as Genetic algorithmic and Rough Set Incorporated Neural Fuzzy System (GARSINFIS) that utilizes the disease rule base auto derived by our projected Genetic rule primarily based Rough Set cluster (GARSC) technique. We have a tendency to mix the benefits of the people and alleviate certain limitation, by fusing numerous soft computing techniques along. Hospital knowledge should be collected as world knowledge for applying GARSINFIS, two established medical knowledge sets square measure benchmarked against different established models that concentrate on the compactness of the derived reasoning rules. All experimental results square measure encouraging, particularly gonad cancer diagnoses. As a result GARSINFIS needs restricted variety of constraints and management parameters. It wants no human intervention and knowledgeable steering to attain correct diagnoses once benchmarked against different models. most vital, it automatically derives rules and choose options that square measure applied and prompt by doctors.

Table III Comparison of Existing Techniques

Author	Year	Model	Processing techniques	Application
Sean N. Ghazavi,[9]	2008	Fuzzy modeling	Fuzzy k-nearest neighbor, fuzzy clustering-based modeling, and fuzzy inference system	Diagnosis of breast cancer dataset
Murat Karabataka [10]	2009	Association Rules	Feature selection method based on Association Rules (AR) and Neural Network (NN)	Diagnosis of erythemato-squamous diseases
A. Zibakhsh [11]	2013	Memetic algorithm with a multi-view fitness function	Evaluates each single fuzzy if–then rule according tothe specified rule quality	Cancer tumor detection
Muhammad Waqar Aslama [12]	2013	Genetic programming	Features selection using t-test, F-score selection, and genetic programming	Diabetes classification
Di Wanga [13]	2014	Hybrid intelligent system	Self-organizing neural fuzzy inference system	Ovarian cancer diagnosis

IV. SUPPORT VECTOR MACHINE (SVM)

Support Vector Machine is constructed on the structural risk minimization principle to seek a decision surface that may separate the data points into two categories with a maximal margin between them. The selection of the correct kernel function is the main challenge when using a SVM. It might have completely different forms like Radial Basis function (RBF) kernel and polynomial kernel. The advantage of the SVM is its capability of learning in sparse, high dimensional spaces with only a few training examples by minimizing the empirical error and the complexness of the classifier at same time.

V. PROPOSED WORK

Genetic algorithms can generate both optimal feature set and SVM parameters at the time. Our analysis objective is to optimize the parameters and feature set, without any lose in the SVM classification accuracy. The proposed technique performs feature selection and parameters setting in an evolutionary way. Feature set selection algorithms are often classified into two categories: the filter approach and the wrapper approach [14]. The wrapper approach to feature set selection is used in paper because of accuracy. Within the literature, few algorithms are proposed for SVM feature selection [15]. Some other GA-based feature selection strategies were also proposed [16]. However, these papers only focused on feature selection, not on the parameters optimization for the SVM classifier. [17] Proposed a GA-based feature selection approach which used theoretical bounds on the generalization error for SVMs.

A. Proposed Algorithm

```
Step 1: Start
Step 2: Read cancer dataset.
Step 3: Select best feature from large attributes of dataset.
Step 4: Set the number of desired features.
Step 5: Set the fitness function Biogafit.
Step 6: call the Genetic Algorithm
Step 6.1: Construction of the first generation
Step 6.2: Selection
   While stopping criteria not met do
Step 6.3: Crossover
Step 6.4: Mutation
Step 6.5: Selection
     End
Step 7: Apply SVM for train and classification
Step 7.1: Loading best feature data from file.
Step 7.2: Initializing and generating Support vector using SVM.
Step 7.3: Apply training to network.
Step 7.5: Testing data against trained network.
Step 8: Calculation of error and accuracy
```

VI. CONCLUSION

Cancer classification, prediction and diagnosis is an emerging research area in the field of Bio-informatics. In this survey various soft-computing methods and machine learning based algorithms for gene selection and cancer classification were discussed in detail. And also we have attempted to explain compare and performing of soft-computing methods which are using of cancer classification , prediction and prognosis to detect it in a earlier stage. specifically in a personalized way we identified a number of trends with respect to the types of computational intelligent methods being used and the types of training data being incorporated ,the kinds of endpoint predictions being made ,the types of cancers being studied, and the overall performance of these methods to predict cancer. Feature selection technique is used to improve accuracy of classifier, reduce dataset and remove irrelevant data. This work gives comparative analysis of various existing feature selection methods and algorithms.

In future better neural network techniques can be incorporated with the present research work for less complexity and better learning adaptability. Moreover, better neuro fuzzy techniques could also be used to improve the classification rate and accuracy.

REFERENCES

[1] Ntzani, E., Ioannidis, J.P., 2003. Predictive ability of DNA microarrays for cancer outcomes and correlates and empirical assessment. Lancet 1 362 (9394) 1439–1444.

[2] Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, Berthold F., Schwab, F.M., Antonescu, C.R., Peterson, C., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673–679

[3] Wang, Y., et al., 2005. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics 21 (8), 1530–1537

[4] Buturovic, L.J, 2006. PCP: a program for supervised classification of gene expression profiles. Bioinformatics 22 (2), 245–247

[5] Ein-Dor, L., et al., 2005. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21 (2), 171–178

[6] Yeung, K.Y., Bumgarner, R.E., Raftery, A.E, 2005. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 21 (10), 2394–2402

[7] Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, Sh, 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21 (5), 631–643

[8] Megha Gupta,Naveen Agarwal "Classification Techniques Analysis", NCCI 2010 -National Conference on Computational Instrumentation

[9] Sean N. Ghazavi, Thunshun W. Liao, ―Medical data mining by fuzzy modeling with selected features‖, Artificial Intelligence in Medicine 43, Pages 195—206, Elsevier, 2008.

[10] Murat Karabataka, M. Cevdet Ince, ―A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases‖, Expert Systems with Applications, 36 Pages 12500–12505, Elsevier, 2009

[11] A. Zibakhsh, M. Saniee Abadeh, ―Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function‖, Engineering Applications of Artificial Intelligence 26, Pages 1274–1281, Elsevier, 2013.

[12] Muhammad Waqar Aslama, Zhechen Zhu, Asoke Kumar Nandi, ―Feature generation using genetic programming with comparative partner selection for diabetes classification‖, Expert Systems with Applications 40, Pages 5402–5412, Elsevier, 2013

[13] Di Wanga, Chai Queka, Geok See Ng, ―Ovarian cancer diagnosis using a hybrid intelligent system with simple yet convincing rules‖, Applied Soft Computing 20, Pages 25–39, Elsevier, 2014.

[14] Kohavi, R., & John, G. (1997). Wrappers for feature subset selection.Artificial Intelligence, 97(1–2), 273–324.

[15] Mao, K. Z. (2004). Feature subset selection for support vector machines through discriminative function pruning analysis. IEEE Transactions on Systems, Man, and Cybernetics, 34(1), 60–67

[16] Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L. A., & Jain, A. K. (2000). Dimensionality reduction using genetic algorithms. IEEE Transactions on Evolutionary Computation, 4(2), 164–171.

[17] Fro¨hlich, H., & Chapelle, O. (2003). Feature selection for support vector machines by means of genetic algorithms. Proceedings of the 15th IEEE international conference on tools with artificial intelligence, Sacramento, CA, US App. 142–148.