ISSN (Print): 2320-5504
ISSN (Online): 2347-4793

Asia Pacific Journal of Research
A peer reviewed international Journal
IMPACT FACTOR : 6.58
Print -ISSN -2320 - 5504
Online-E-ISSN - 2347 - 4793
University Grants Commission
Approved Journal S.N.45797

PERFORMANCE ANALYSIS OF MACHINE LEARNING TECHNIQUES FOR MICROARRAY GENE EXPRESSION CANCER DIAGNOSIS

Dr.S.Sasikala

Head, Department of Computer Science

Sree Saraswathi Thyagaraja College, Pollachi, Coimbatore, Tamil Nadu, India

ABSTRACT

The multicategory cancer classification is playing a vital role in the field of medical sciences. As the numbers of cancer victims are increasing steadily, the necessity of the cancer classification techniques has become indispensible. Effective cancer classification is very important for the cancer diagnosis and it is one of the active areas of research in the field of medical science. This research mainly focuses on the development of an effective multicategory classification for microarray gene expression cancer diagnosis using machine learning techniques for significant accuracy, reliability and less error rate.

Key words: cancer Classification, Multicategory, Microarray, Gene expression, Machine Learning

I. INTRODUCTION

Fundamentally cancer [1] is described by an abnormal, uncontrolled growth that may demolish and invade neighboring healthy body tissues or elsewhere in the body. Our body is composed of many millions of tiny cells, each a self-contained living unit. Normally, each cell coordinates with the others that compose tissues and organs of our body. The human body consists of billions of cells, majority of the cells include a restricted life-span and is being replaced cyclic manner. Every cell is competent of duplicating themselves. In rare situation there is some fault in the division and a rogue, potentially malignant cell arises. Normal cells in the body grow and divide for a period of time and then stop growing and dividing. Thereafter, they only reproduce themselves as necessary to replace defective or dying cells. Cancer occurs when this cellular reproduction process goes out of control. In other words, cancer is a disease characterized by uncontrolled, uncoordinated and undesirable cell division. Unlike normal cells, cancer cells continue to grow and divide for their whole lives, replicating into more and more harmful cells.

II. MACHINE LEARNING

Machine learning is a scientific discipline that is concerned 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. A 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 main aim of a learner is to generalize from its past experience. The training data from its experience come from some generally unknown probability distribution and the learner has to discover something more general, something about that distribution that facilitates it to produce effective responses for the future cases**.**

III. ANALYTIC NETWORK PROCESS (ANP)

ANP is a common type of Analytical Hierarchical Analysis (AHP). Saaty [[2], [3], [4]] suggested the use of AHP to handle the problem of independence on alternatives or criteria, and the use of ANP to solve the problem of dependence among alternatives.AHP approaches form a structure of the decisions that employs a one-way hierarchical relation with regard to decision layers.

ANP was also introduced by Saaty. It is a generalization of the AHP [5]. The main difference between AHP and ANP is that, AHP represents a framework with a unidirectional hierarchical relationship but ANP is developed for the subjective evaluation of a group of

ISSN (Print): 2320-5504
ISSN (Online): 2347-4793

alternatives according to multiple criteria organized in a hierarchical structure.

The top element of the hierarchy in AHP is usually the overall goal for the decision model. The hierarchy decomposes the common to more particular properties until a level of manageable decision criteria is obtained. ANP does not need this hierarchical structure; it facilitates factors to 'control' and be 'controlled' by the varying levels or `clusters' of attributes. Some controlling factors are also present at the same level [6].

This interdependency among factors and their levels is defined as a 'systems with feedback' technique. AHP does not consist of feedback loops among the factors that can regulate weightings and reduce the opportunity of the reverse ranking technique. The relative significance of the impacts on a given element is calculated on a ratio scale related to AHP.ANP facilitates for complex interrelationships among decision levels and properties. ANP feedback technique replaces hierarchies with networks in which the relationships between levels are not easily denoted as higher or lower, dominated or being dominated, directly or indirectly [7].

For example, not only does the significance of the criteria decide the importance of the alternatives as in a hierarchy, but also the importance of the alternatives may have an impact on the importance of the criteria [STY80]. Thus, a hierarchical structure with a linear top-to-down form is not appropriate for a complex system. ANP approach is competent to deal with interdependent relationships among the elements by acquiring the composite weights through the development of a super matrix. The super matrix concept contains parallels to the Markov chain process [8], where relative importance weights are altered by forming a super matrix from the eigenvectors of these relative importance weights. The weights are, then, altered by deciding products of the super matrix.

IV. STATISTICAL RANKING TECHNIQUES

ANOVA

Analysis of variance (ANOVA) is a parametric statistical approach used to evaluate datasets. It is comparable in application to techniques like T-Test and Z-Test, in that it is used to evaluate means and the relative variance among them, even though ANOVA is best applied where more than 2 populations or samples are meant to be evaluated.

Using ANOVA over other methods like t-tests provides considerable advantages such as:

- ANOVA lessens the probability of a type-I error. Making multiple assessments increase the likelihood of discovering something by chance—making a type-I error.
- It has more than two experimental conditions. The T-Test for correlated means can only evaluate two conditions.
- A larger sample size and a larger amount of samples cannot be evaluated using t-test.
- T-Test cannot be used to assess much bigger and more difficult problems, by means of multiple variables and datasets.
- Further there is an added benefit of ANOVA when compared with simple t-tests and ANOVA is that it is very simple to identify interaction consequences among variables.

So, this work uses the ANOVA ranking technique for preprocessing the genes.

V. METHODOLOGY

The proposed methodology uses the machine learning techniques for developing an effective cancer classification system. The methodology aims at developing an efficient multicategory classification for microarray gene expression cancer diagnosis which provides reliable results. The methodologies used in the proposed approach are

Phase I: Effective Multicategory Classification Using ELM-ANP Approach for Microarray Gene Expression Cancer Diagnosis

This approach uses the Support Vector Machine (SVM) for developing an effective cancer classification system. This approach deals with the advanced and developed methodology know for cancer multi classification using Support Vector Machine (SVM)-Analytic Network Process (ANP) for microarray gene expression cancer diagnosis, this is used for directing multicategory classification problems in the cancer diagnosis area. ANP are decision making approach for solving the problem of independence on alternatives or criteria. ANP techniques form a framework of the decisions that uses a one-way hierarchical relation with respect to decision layers. In this, the weight matrix for the SVM is obtained using ANP approach. SVMs are an appropriate new technique for binary classification tasks, which is related to and contain elements of non-parametric applied statistics, neural networks and machine learning. SVMs can generate accurate and robust classification results on a sound theoretical basis, even when input data are non-monotone and non-linearly separable. The performance of SVM is evaluated for the multicategory classification on benchmark microarray data sets for cancer diagnosis, namely, Lymphoma, Leukemia and SRBCT Data set. The results indicate that SVM produces comparable or better classification accuracies when the data given as input are preprocessed.

Phase II: Effective Multicategory Classification Using ELM-ANP Approach for Microarray Gene Expression Cancer Diagnosis

In this approach, a new kind of ELM called Fast ELM is used. ANOVA ranking is used for ranking the best genes in the data sets. Then the genes are classified using ELM. ELM is trained using Levenberg-Marquat algorithm. Fast ELM avoids problems such as local minima; improper learning rate and over fitting usually happens in iterative learning techniques and finishes the training very fast. ANP is integrated with the Fast ELM approach for finding the weight factor w for the ELM. Then the genes are classified using the Fast ELM algorithm which integrates ANP approach.

ISSN (Print): 2320-5504
ISSN (Online): 2347-4793

Phase III: Improved Multicategory Classification Using RVM-ANP Approach for Microarray Gene Expression Cancer Diagnosis

This approach deals with the advancement in cancer multicategory classification using Relevance Vector Machine (RVM) and ANP for microarray gene expression cancer diagnosis. The proposed technique can be highly used for directing multicategory classification problems in the cancer diagnosis area. SVM and ELM techniques are used for binary classification tasks, which is related to and contains elements of non-parametric applied statistics, neural networks and machine learning. The cancer classification using the present approach does not provide the expected accuracy and sometimes the result of clustering may be wrong. To overcome this problem an efficient cancer classification using the RVM-ANP used. Initially, ANOVA ranking is used for ranking the genes. Then the ranked genes are given as input to the ANP approach for finding the weighted factor the RVM. Then the RVM algorithm is used for classification. This learning algorithm can generate accurate and robust classification results on a sound theoretical basis, even when input data are non-monotone and non-linearly separable.

In order to evaluate the performance of the proposed approaches for multicategory cancer diagnosis three datasets are used. They are SRBCT, Lymphoma and Leukemia. The parameters used for the evaluation of the proposed approaches are:

- Training Accuracy
- Training Time

VI. OBSERVATIONS AND FINDINGS

A detailed discussion of the performance of the SVM-ANP, RVMANP, ELM-ANP techniques is presented.

i) *Results of Lymphoma Datasets*

The evaluation of the testing accuracy and the training time of the proposed approaches for the Lymphoma are presented in the following section.

A. Testing Accuracy

TABLE: 1
ACCURACY COMPARISON FOR THE PROPOSED APPROACHES

S. No	No. of Gene Combination	5-Fold CV			10-fold CV		
		SVM with ANP	ELM with ANP	RVM with ANP	SVM with ANP	ELM with ANP	RVM with ANP
1	100,2	91.25	92.34	96.34	89.11	91.47	96.41
2	100,3	92.14	93.15	97.64	90.78	92.62	97.53

The average testing accuracy of the SVM with ANP, ELM with ANP and RVM with ANP are observed in Table: 1. It is clearly observed from the Table: 1 that the proposed RVM approach provides very high testing accuracy when compared to SVM with ANP and ELM with ANP approaches.

B. Training Time

The average training time taken by the three proposed cancer classification techniques for the lymphoma dataset is compared Table: 2. It clearly shows that the proposed RVM with ANP cancer classification approach is processed in very less time when comparing with the other two approaches.

TABLE: 2
TRAINING TIME COMPARISON OF THE PROPOSED APPROACHES

S. No	No. of Gene Combination	Training Time (sec)		
		SVM with ANP	ELM with ANP	RVM with ANP
1	100,2	21.985	4.444	3.894
2	100,3	18.021	5.015	4.982

7.1.1. Results of Leukemia Dataset

The evaluation of the testing accuracy and the training time of the proposed approaches for the Leukemia are presented in the following section.

A. Testing Accuracy

TABLE :3
TESTING ACCURACY COMPARISON FOR THE PROPOSED APPROACHES

S. No	No. of Gene Combination	5-Fold CV			10-fold CV		
		SVM with ANP	ELM with ANP	RVM with ANP	SVM with ANP	ELM with ANP	RVM with ANP
1	100,2	91.34	92.66	96.75	90.01	92.41	96.14
2	100,3	91.67	92.86	97.66	90.10	92.93	97.16

Table :3 shows the average testing accuracy of the proposed SVM with ANP, ELM with ANP and RVM with ANP approaches. For both the 2-Gene and 3-Gene combinations, the proposed RVM with ANP approach is observed to produce better and effective results.

B. Training Time

The average training time taken by the three proposed cancer classification techniques for the lymphoma dataset is compared Table :4. It clearly shows that the proposed ELM with ANP cancer classification approach is processed in very less time when comparing with the other two approaches.

TABLE : 4
TRAINING TIME COMPARISON OF THE PROPOSED APPROACHES

S. No	No.of Gene Combination	Training Time (sec)		
		SVM with ANP	ELM with ANP	RVM with ANP
1	100,2	21.985	4.444	3.894
2	100,3	18.021	5.015	4.982

ISSN (Print): 2320-5504
ISSN (Online): 2347-4793

7.1.2. Results of SRBCT Dataset

The performance evaluation based on the testing accuracy and the training time of the proposed approaches for the Lymphoma are presented in the following section.

A. Testing Accuracy

Table: 5 show the testing accuracy of the proposed classification approaches. The results for both the 2-gene and 3-gene combinations are presented. The testing accuracy of the proposed RVM with ANP approach for the 5-fold and 10-fold CV is observed to be very high when compared with the other two proposed approaches.

TABLE : 5
TESTING ACCURACY COMPARISON FOR THE PROPOSED APPROACHES

S.No	No. of Gene Combination	5-Fold CV			10-fold CV		
		SVM with ANP	ELM with ANP	RVM with ANP	SVM with ANP	ELM with ANP	RVM with ANP
1	100,2	91.01	93.36	96.78	90	93.72	96.74
2	100,3	91.98	94.98	97.85	91.17	94.93	97.36

B. Training Time

The average training time taken by the proposed techniques for the SRBCT dataset is compared in Table : 6. It clearly shows that the proposed ELM approach train the system in very less time when comparing with the other two approaches.

TABLE : 6
TRAINING TIME COMPARISON OF THE PROPOSED APPROACHES

S.No	No.of Gene Combination	Training Time (sec)		
		SVM with ANP	ELM with ANP	RVM with ANP
1	100,2	41.23	3.124	4.568
2	100,3	43.45	3.985	5.214

7.2. OVERALL RANKING BASED ON PERFORMANCE

It is clear from the experimental results that the performance of the proposed approaches are better in terms of testing accuracy and training time than the standard approaches like SVM, ELM and RVM. The following section provides the ranking of the proposed approaches based on the above mentioned parameters for three datasets.

Table : 7 shows the overall ranking performance of the proposed approached. This overall ranking is based on the experimental results of all the performance parameters. It is observed from the table that the third proposed RVM with ANP is ranked first in the overall testing accuracy performance as it outperforms better the other approaches.

TABLE: 7
OVERALL TESTING ACCURACY RANKING PERFORMANCE OF THE CLUSTERING APPROACHES FOR ALL DATASETS

Proposed Approaches	Lymphoma	Leukemia	SRBCT
SVM with ANP	3	3	3
ELM with ANP	2	2	2
RVM WITH ANP	1	1	1

TABLE: 8
OVERALL TESTING ACCURACY RANKING PERFORMANCE OF THE CLUSTERING APPROACHES FOR ALL DATASETS

Proposed Approaches	Lymphoma	Leukemia	SRBCT
SVM with ANP	3	3	3
ELM with ANP	2	1	1
RVM WITH ANP	1	2	2

Table: 8 shows the overall training time of the proposed approaches. It is observed from the table that, for Leukemia and SRBCT data sets, the ELM with ANP approach is observed to produce better training time. In lymphoma dataset, RVM with ANP is observed to produce better results in terms of training time.

VII. CONCLUSION

From the experimental results, it can be clearly observed that the proposed techniques results performance than the other standard approaches. Moreover, the proposed RVM with ANP shows higher testing accuracy than other proposed techniques like SVM with ANP and ELM with ANP.

In terms of training time, it is observed that, the proposed ELM with ANP trains the data in very less time when compared other proposed approaches. This is mainly due to the Levenberg-Marquat algorithm.

From the experimental results, it is found that the proposed RVM with ANP is very effective in terms of prediction Accuracy. The proposed ELM with ANP is very effective in terms of the training time. SVM is best when its complexity is taken into consideration.

This paper focuses on the approach for developing efficient cancer detection systems using machine learning techniques. In this paper, machine learning techniques like SVM, Fast ELM and RVM is used for cancer classification.

The first approach uses the SVM technique for classification. In the preprocessing step, ANOVA ranking is used for ranking the genes. In order to improve the performance of the cancer classification system, SVM is integrated with the Analytic Network Process (ANP). The ANP approach is capable of handling interdependent

ISSN (Print): 2320-5504
ISSN (Online): 2347-4793

relationships among the elements by obtaining the composite weights through the development of a supermatrix. Thus the weights calculated from the ANP are applied to the SVM classifier and then the genes are classified using the SVM classifier.

The second approach used the ELM classification technique. Levenberg Marquat algorithm is used for training the ELM in lesser time. Then ANP is integrated with ELM for better performance. Similarly, third approach uses the RVM with ANP cancer classification technique.

The experiments are evaluated in three datasets namely Lymphoma, Leukemia and SRBCT. The performances of the three approaches are evaluated based on the parameters such as Testing Accuracy and Training Time. From the experimental results, it is observed that, the RVM with ANP approach provides best testing accuracy results for all the datasets considered. In terms of training time, ELM with ANP is observed to take lesser training time for Leukemia and SRBCT datasets.

Thus it is clear that, the "Improved Multicategory Classification Using RVM-ANP Approach for Microarray Gene Expression Cancer Diagnosis" is very efficient in cancer classification with very high accuracy. Similarly, the "Effective Multicategory Classification Using ELM-ANP Approach For Microarray Gene Expression Cancer Diagnosis" approach provides very effective training time results and provides classification of genes in lesser time.

VIII. REFERENCES

1. Anand P, Kunnumakkara AB and Kunnumakara AB, "Cancer is a preventable disease that requires major lifestyle changes", Pharm. Res., Vol. 25, No. 9, 2008.
2. Saaty, T.L., "Interaction and Impacts in Hierarchical Systems", Academic Press, Inc., Pp.29–77, 1980.
3. Saaty, T.L., "The Analytic Hierarchy Process", New York. McGraw-Hill Book Company, 1980.
4. [4] Saaty, T.L., "Fundamentals of the Analytic Hierarchy Process", RWS Publications, Pittsburgh, PA, 2006.
5. T.L Saaty, "Decision Making with Dependence and Feedback-The Analytic Network Process", RWS Publications, Pittsburgh, 1996.
6. Babak Daneshvar Rouyendegh and Serpil Erol, "The DEA – FUZZY ANP Department Ranking Model Applied in Iran Amirkabir University", Acta Polytechnica Hungarica, Vol. 7, No. 4, 2010.
7. L. M. Meade, and J. Sarkis, "Analyzing Organizational Project Alternatives for Agile Manufacturing Processes: An Analytical Network Approach", International Journal of Production Research, Vol. 37, Pp. 241-261, 1999.
8. Saaty, T.L., "The Analytic Hierarchy Process", New York. McGraw-Hill Book Company, 1980.