International Journal of Biosciences | IJB |
ISSN: 2220-6655 (Print), 2222-5234 (Online)
http://www.innspub.net
Vol. 20, No. 5, p. 169-175, 2022

INNSPUB

RESEARCH PAPER OPEN ACCESS

Extreme learning machine for cancer classification from miRNA gene expression data

Ansuman Kumar*, Anindya Halder

Department of Computer Application, North-Eastern Hill University, Tura Campus, Meghalaya - 794002, India

Key words: Extreme learning machine, Cancer classification, miRNA gene expression data.

http://dx.doi.org/10.12692/ijb/20.5.169-175 Article published on May 28, 2022

Abstract

Cancer classification from microRNA (miRNA) gene expression data is a difficult task in system biology and machine learning as conventional classification methods require a sufficiently large number of labeled samples to train the classifiers accurately, particularly when the labeled samples are very expensive and difficult to collect. Therefore, conventional classification methods usually do not provide the desired classification accuracy due to the scarcity of training samples. In this context, we present an extreme learning machine (ELM) technique for cancer classification from miRNA gene expression data that can improve the classification accuracy as it is extremely fast and accurate compared to other traditional methods. The presented method is evaluated using publicly available miRNA gene expression datasets of breast cancer, pancreatic cancer, colorectal cancer, prostate cancer and lung cancer in terms of classification accuracy, precision, recall, macro F_1-measure, micro F_1-measure and kappa in comparison to four other state-of-the-art methods. Experimental results justify the dominance of the ELM method over the other compared methods for cancer sample classification from miRNA Gene Expression data.

* **Corresponding Author:** Ansuman Kumar ✉ ansuman.kumar@gmail.com

Introduction

Cancer is one of the dangerous diseases due to the unusual rapid division and unregulated growth of the cells (Kumar *et al.,* 2020). It is one of the leading causes of death across the globe. There was approximately 18.1 million new cancer patients and 9.9 million cancer-related deaths worldwide reported in 2020 (Sung *et al.,* 2021). The number of new cancer patients per year is expected to increase by 29.5 million and the number of cancer-related deaths approaching 16.4 million by the year (Sung *et al.,* 2021). Therefore, early detection and diagnosis of cancer have become an essential area of research for biologists and researchers across the world. In this context, it is necessary to construct an accurate and reliable classifier that can be used by the physicians to discriminate benign tumors from malignant tumors without going for a surgical biopsy (Marak *et al.,* 2021). Surgical biopsy tests are extremely invasive as tissue samples are needed to be extracted from patients in the form of proteins. Although, conventional protein-based diagnostic methods require careful analysis as well as it produces a less accurate result. However, recent researches have emphasized the role of non-protein-coding ribonucleic acid (ncRNA) in cancer (Esquela-Kerscher *et al.,* 2006). microRNA (miRNA) is one of type of ncRNAs that handles proliferation, differentiation, development, and apoptosis (Hwang *et al.,* 2006). It is a small, single-stranded, non-coding endogenous RNAs of approximately 22 nucleotides (nt) length that manage gene expression by controlling their target mRNAs for translation repression. The miRNA expression levels differ significantly between cancerous and non-cancerous cells that recommend that miRNAs might be involved in the development of cancer and may even be used in the diagnosis and treatment of cancer (Marak *et al.,* 2021). Several machine learning methods have been applied in classifying tumors using gene expression data (Pirooznia *et al.,* 2008, Tarek *et al.,* 2017). These methods can broadly be classified as supervised (Haider *et al.,* 2013; Vanitha *et al.*, 2015), semi-supervised (Marak *et al.,* 2021; Halder *et al.,* 2014), active learning (Kumar *et al.,* 2019; Halder *et al.,* 2019), and ensemble based (Kumar *et al.,* 2020; Chen *et al.,* 2012) methods etc. Classification of miRNA gene expression data usually depends on traditional supervised methods that require sufficient number of manually labeled training samples to predict unlabeled samples to a particular class. Although miRNA gene expression labeled samples are expensive, time-consuming, and challenging to collect, whereas unlabeled samples are relatively inexpensive and easy to gather. Therefore, the limited training samples are a bottleneck to be used in traditional supervised methods for cancer classification. In this context, it is a challenging to construct a robust classifier that can produce high accuracy in classifying cancerous samples from miRNA gene expression data. Motivated from the above said challenges, an extreme learning machine (ELM) is used in this article, which is extremely fast compared to other traditional methods as it is implemented without iteration and no human-intervention is needed. The advantage of ELM over other neural network algorithms (i.e., backpropagation (BP) based algorithm) is that the learning parameters of hidden nodes, input weights and biases are randomly assigned and need not be tuned and the output weights can be analytically computed by the simple generalized inverse operation (Ding *et al.*, 2013; Huang *et al.*, 2015).

The ELM method is evaluated using publicly available miRNA gene expression datasets (Clough *et al.*, 2016) of pancreatic cancer, colorectal cancer, prostate cancer, lung cancer and breast cancer in terms of six validity measures viz., percentage accuracy, precision, recall, macro F_1, micro F_1, and kappa. The classification performance of the ELM method is compared with three other state-of-the-art methods namely, *k*-nearest neighbour (KNN) classifier (Aha *et al.*, 1991), support vector machine (SVM) classifier (Vanitha *et al.* 2015) and Naïve Bayes (NB) classifier (Chandra *et al.*, 2011). The overall results reveal that employing the extreme learning machine classifier in miRNA gene expression data can achieve better accuracy. The rest of the article is organised as follows. In Section 2, we provide material and methods. Experimental results and discussions are reported in Section 3. Finally, conclusions and future direction of research are highlighted in Section 4.

Material and methods

The extreme learning machine (ELM) method is used for cancer classification from miRNA Gene Expression data. Thus, brief description of extreme learning machine is highlighted here. Datasets used for the experiment along with the brief description of the other compared methods are also reported followed by performance evaluation metrics at the end of this section.

Extreme learning machine

Extreme Learning Machine (ELM) was introduced by Huang *et al.*, (Huang *et al.,* 2006). It is feedforward neural networks having a single hidden layer. Parameters of hidden layer are assigned randomly and need not be tuned in learning process. Input layer weights *w* and biases *b* are also assigned randomly and never adjust them due to the input weights are fixed in ELM method. The output weights *β* are independent of them (unlike in the backpropagation training method) and have a straight forward solution that does not require iteration (Akusok *et al.*, 2015). Therefore, ELM method computes linear output layer very fast compared to backpropagation networks. The block diagram of ELM method is shown in Fig. 1.

$$y_j = \sum_{i=1}^{n_h} \beta_i g\left(\omega_i . x_j + b_i\right), \quad j = 1, \dots, N. \qquad (1)$$

where n_h is a number of hidden neurons, N is a number of training samples, *g(.)* is an activation function, $\boldsymbol{w_i}$ is the weight vector connecting the i^{th} hidden neuron to the input layer, β_i is the output weight vector connecting the i^{th} hidden neuron to the output layer, b_i is the bias of the i^{th} hidden neuron, and $\boldsymbol{w_i}$ $\boldsymbol{.x_j}$ is the inner product of $\boldsymbol{w_i}$ and $\boldsymbol{x_j}$. We can shorten the Equation (1) by taking $g\left(\omega_i . x_j + b_i\right)$ as H and rewrite the equation as follows:

$$Y = H\beta, \qquad (2)$$

where,

$$H = \begin{bmatrix} g(\omega_1 . x_1 + b_1) & \cdots & g\left(\omega_{n_h} . x_1 + b_{n_h}\right) \\ \vdots & \ddots & \vdots \\ g(\omega_1 . x_N + b_1) & \cdots & g\left(\omega_{n_h} . x_N + b_{n_h}\right) \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \vdots \\ \beta_{n_h}^T \end{bmatrix}, Y = \begin{bmatrix} y_1^T \\ \vdots \\ y_N^T \end{bmatrix},$$

and *g(.)* is a non-linear piecewise continuous function such as Sigmoid function or Gaussian function (Huang *et al.,* 2006). The output weight *β* is computed based on the labeled target *Y* as follows:

$$\beta = (H^T H)^{-1} H^T Y = H^{\dagger} Y, \qquad (3)$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse (Akusok *et al.*, 2015) of the hidden layer output matrix *H*.

The datasets

The experiments are carried out on eight miRNA gene expression datasets (viz., GSE24279, GSE85589, GSE30454, GSE60117, GSE102286, GSE51853, GSE26659 and GSE58606) of five cancer types namely, pancreatic, colorectal, prostate, lung and breast cancers.

These datasets are downloaded from the Gene Expression Omnibus (GEO) (Clough *et al.*, 2016). Each miRNA dataset is uniquely identified by the accession ID. The datasets comprise of non-cancerous and cancerous samples, and each sample consists of gene expression values along with class label information. The summary of each dataset, such as the cancer type, accession ID, total number of samples, number of cancerous samples, number of non-cancerous samples and number of genes in each sample are provided in Table 1. Detailed descriptions of the used datasets are given below:

Pancreatic cancer

GSE24279 and GSE85589 pancreatic cancer miRNA datasets are used for the experiments. GSE24279 dataset consists of 158 samples (136 cancerous and 22 non-cancerous samples) with each sample containing 848 miRNAs gene expression values. GSE85589 dataset comprises of 88 cancerous and 19 non-cancerous samples with each sample containing 2579 miRNAs gene expression values.

Colorectal cancer

GSE30454 colorectal cancer miRNA dataset is used for the experiments. This dataset contains 74 samples out of which 20 samples are cancerous and 54 samples are non-cancerous and each sample is having 1145 genes.

Prostate cancer

GSE60117 prostate cancer miRNA dataset consists of 77 samples out of which 56 samples are cancerous and 21 samples are non-cancerous and each sample is described by 2689 miRNAs gene expression values.

Lung cancer

Two miRNA datasets (GSE102286 and GSE51853) of lung cancer are used for the experiments. GSE102286 dataset contains 179 observations in which 88 samples are of cancerous and 91 samples are of non-cancerous and 734 expression values present per sample. GSE51853 dataset comprises of 131 samples in which 126 samples are cancerous and 5 samples are non-cancerous and each sample is measured over 470 genes.

Breast cancer

Two miRNA datasets (GSE26659 and GSE58606) of breast cancer are used for the experiments. These datasets are briefly described as follows.

GSE26659 dataset is having 94 samples out of which 17 samples are cancerous and 77 samples are non-cancerous. Each sample is described by 237 gene expression values. GSE58606 dataset consists of 1926 gene expression values for each sample and it comprises of 122 cancerous and 11 non-cancerous samples.

The compared methods

We compared the performance of the ELM method (in terms of all the validity metrics) with respect to three other state-of-the-art methods namely, *k*-nearest neighbour (KNN) classifier (Aha *et al.*, 1991), support vector machine (SVM) classifier (Vanitha *et al.* 2015) and Naïve Bayes (NB) classifier (Chandra *et al.*, 2011). The brief descriptions of kNN, SVM and NB methods are as follows.

k-nearest neighbour (KNN) is the simplest method for classification. In this method, class label of the test sample is assigned based on the *k*-nearest neighbours labeled samples of that test sample (Aha *et al.*, 1991), where *k* is the positive number.

Support vector machine (SVM) is a supervised machine learning technique that can be used for classification as well as regression problems under statistical techniques. It handles non-linear decision boundaries of arbitrary complexity (Vanitha *et al.*, 2015).

The decision boundary (a straight line in the case of a two-dimensional separation) is positioned to leave the largest possible margin on either side. Classification is done by the finding the hyper-plane that differentiates the two classes very well.

Naïve Bayes classifier (Chandra *et al.*, 2011) is also supervised learning algorithm. It is based on Bayes theorem and used for solving classification problems. Naïve Bayes classifier is one of the simple and most effective classification algorithms which helps in making the machine learning models that can make fast predictions.

Performance validity metrics

Six different kinds of validity metrics (viz., percentage accuracy, precision, recall, macro averaged F_1, micro averaged F_1 (Kumar *et al.*, 2019), and kappa (Cohen, 1960) are used to assess the performance of the all the methods.

Results and discussion

In this article, we have reported the average results of 10 simulation runs of all the methods performed on eight real life microarray gene expression datasets. The ELM method is implemented in MATLAB and the other three methods, KNN, SVM and NB are simulated using WEKA 3.8.3 (Waikato Environment for Knowledge Analysis) tool in 64-bit Windows 10 machine with processor speed 2.50 GHz and 4 GB RAM. The experiments are carried out with the same number of training samples, i.e, 20% of the total samples for all the methods (viz., KNN, SVM, NB, and ELM).

Table 1. Summary of the eight-miRNA gene expression cancer datasets used for the experiments.

Cancer Type	Accession ID	# Total Samples	# Cancerous Samples	# Non-cancerous Samples	# Genes /Sample
Pancreatic Cancer	GSE24279	158	136	22	848
	GSE85589	107	88	19	2579
Colorectal Cancer	GSE30454	74	20	54	1145
Prostate Cancer	GSE60117	77	56	21	2689
Lung Cancer	GSE102286	179	88	91	734
	GSE51853	131	126	5	470
Breast Cancer	GSE26659	94	17	77	237
	GSE58606	133	122	11	1926

The summary of the average experimental results of 10 simulations on eight miRNA gene expression datasets achieved by the ELM and compared methods in terms of six validity metrics (viz., percentage accuracy, precision, recall, macro F_1, micro F_1, and kappa) are reported in Table 2.

Table 2. Summary of the average experimental results (in terms of accuracy, precision, recall, macro F_1, micro F_1 and kappa) of 10 simulations achieved by different methods viz., KNN, SVM, NB and ELM performed on eight microarray gene expression datasets.

Cancer Type	Accession ID	Methods	Accuracy (%)	Overall Precision	Overall Recall	Macro F_1	Micro F_1	Kappa
Pancreatic Cancer	GSE24279	KNN	87.30±5.20	0.8590	0.8730	0.8260	0.8494	0.2329
		SVM	85.71±5.48	0.7590	0.8570	0.8010	0.8252	0.1429
		NB	86.51±2.90	0.8560	0.9500	0.8210	0.8325	0.1429
		ELM	**94.84**±1.29	**0.9609**	**0.9727**	**0.8278**	**0.8536**	**0.6589**
	GSE85589	KNN	84.88±3.20	0.8320	0.8490	0.8320	0.8445	0.3957
		SVM	82.24±6.12	0.8340	0.8220	0.7900	0.8112	0.2310
		NB	84.55±4.33	0.8598	0.8260	0.7790	0.8088	0.1744
		ELM	**91.49**± 4.73	**0.8910**	**0.9526**	**0.8376**	**0.8634**	**0.6823**
Colorectal Cancer	GSE30454	KNN	81.35±8.21	0.8870	0.8140	0.8220	0.8490	0.6189
		SVM	93.22±4.67	0.9230	0.9320	0.9340	0.9398	0.8455
		NB	72.88±5.30	0.8040	0.7290	0.6300	0.7212	0.0817
		ELM	**95.66**±4.17	**0.9255**	**0.9723**	**0.9427**	**0.9479**	**0.8864**
Prostate Cancer	GSE60117	KNN	85.48±7.12	0.8590	0.8550	0.8400	0.8329	0.5578
		SVM	74.19±9.75	0.7150	0.7420	0.7542	0.7510	0.2581
		NB	83.87±7.37	0.8320	0.8390	0.8310	0.8420	0.5414
		ELM	**95.05**± 6.14	**0.9271**	**0.9655**	**0.9378**	**0.9453**	**0.8780**
Lung Cancer	GSE102286	KNN	90.50±2.90	0.9200	0.9050	0.9040	0.9120	0.8106
		SVM	50.84±9.40	0.5512	0.5080	0.4872	0.4900	0.4916
		NB	85.31±6.18	0.8710	0.8530	0.8520	0.8865	0.7086
		ELM	**93.58**±3.41	**0.9401**	**0.9352**	**0.9353**	**0.9377**	**0.8713**
	GSE51853	KNN	95.85±8.55	0.9422	0.9530	0.9544	0.9550	0.3210
		SVM	95.23±8.44	0.9660	0.9520	0.8860	0.8980	0.2988
		NB	**96.94**±7.65	**0.9700**	**0.9690**	**0.9600**	**0.9590**	0.3247
		ELM	92.85±9.22	0.6634	0.9639	0.6938	0.7796	**0.4190**
Breast Cancer	GSE26659	KNN	97.33±2.10	0.9770	0.9730	0.9740	0.8542	0.9123
		SVM	92.00±3.33	0.9270	0.9200	0.9100	0.8874	0.6586
		NB	93.33±2.78	0.9380	0.9330	0.9270	0.9020	0.7257
		ELM	**98.72**±1.91	**0.9580**	**0.9928**	**0.9724**	**0.9745**	**0.9451**
	GSE58606	KNN	**92.59**±4.66	0.9260	**0.9260**	**0.9260**	0.9210	0.0380
		SVM	91.51±4.90	0.9178	0.9150	0.9002	0.8990	0.2432
		NB	92.45±6.56	**0.9500**	0.9250	0.9330	**0.9394**	**0.6271**
		ELM	90.17±5.05	0.6187	0.9009	0.6536	0.7314	0.3331

The best results obtained for each dataset are marked with bold font in the table and the standard deviations of percentage accuracies of 10 simulations are also shown using ± sign in Table 2.

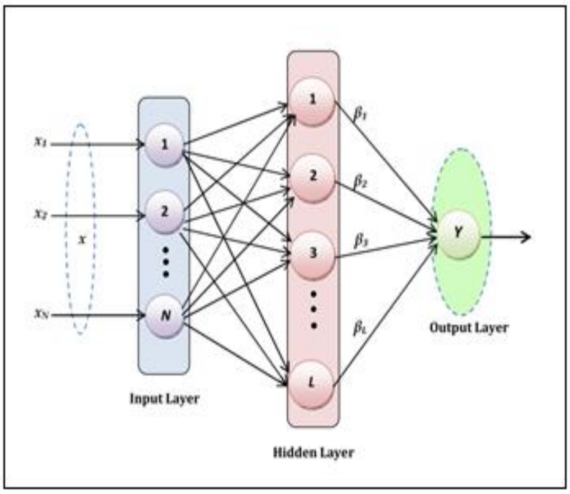

Fig. 1. Block diagram of the Extreme Learning Machine (ELM).

We can observe from the summarized experimental results (Table 2), that the ELM method outperformed the other counter-part methods for six datasets (viz., GSE58606, GSE24279, GSE85589, GSE30454, GSE60117, GSE102286 and GSE26659), whereas in two cases (viz., GSE51853 andGSE58606) other methods NB and KNN respectively performed better in terms of accuracy compared to the ELM method.

Conclusion

Traditional supervised learning methods require a large amount of labeled training data to achieve desired classification accuracy. Therefore, small labeled sample size in miRNA gene expression data remains a bottleneck in obtaining robust and accurate classifier. In order to resolve these issues, we use extreme learning machine (ELM) classifier for cancer sample classification from miRNA gene expression datasets. The efficiency of this method is validated using eight publicly available miRNA gene expression cancer datasets in terms of six different kinds of validity metrics viz., accuracy, precision, recall, macro F_1-measures, micro F_1-measures and kappa. It can be observed from the experimental results that the ELM method dominated the other compared methods in terms of all most all the validity measures (viz., accuracy, overall precision, overall recall, macro averaged F_1 measure, micro averaged F_1 measure and kappa) for six datasets namely, GSE58606, GSE24279, GSE85589, GSE30454, GSE60117, GSE102286 and GSE26659, whereas in two datasets (viz., GSE51853 and GSE58606) other methods NB and KNN respectively performed better in terms of accuracy compared to the ELM method. The encouraging results obtained from the ELM method may motivate researchers to apply this method in other application domains particular where the labeled samples are limited. The ELM method may also be tested on other microarray /miRNA gene expression cancer datasets in future.

References

Aha DW, Kibler D, Albert MK. 1991. Instance-Based Learning Algorithms. Machine Learning **6,** 37–66.
https://doi.org/10.1007/BF00153759

Akusok A, Bjrk K, Miche Y, Lendasse A. 2015. High-performance extreme learning machines: A complete toolbox for big data applications. IEEE Access **3,** 1011–1025.
https://doi.org/10.1109/ACCESS.2015.2450498

Chandra B, Gupta M. 2011. Robust approach for estimating probabilities in naïve Bayesian classifier for gene expression data. Expert Systems with Applications **38(3),** 1293-1298.
https://doi.org/10.1016/j.eswa.2010.06.076

Chen X, Ishwaran H. 2012. Random forests for genomic data analysis. Genomics **99(6),** 323–329.
https://doi.org/10.1016/j.ygeno.2012.04.003

Clough E, Barrett T. 2016. The gene expression omnibus database. Statistical Genomics: Methods in Molecular Biology, Humana Press, New York **1418,** 93–110.
https://doi.org/10.1007/978-1-4939-3578-9_5

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20,** 37–46.
https://doi.org/10.1177/001316446002000104

Ding S, Zhao H, Zhang Y, Xu X, ru N. 2013. Extreme learning machine: algorithm, theory and applications. Artificial Intelligence Review **44(06),** 1–8. https://doi.org/10.1007/s10462-013-9405-z

Esquela-Kerscher E, Slack FJ. 2006. Oncomirs-micro RNAs with a role in cancer. Nature reviews cancer **6(4),** 259–269. https://doi.org/10.1038/nrc1840

Haider AA, Asghar S. 2013. A survey of logic-based classifiers. International Journal of Future Computer and Communication **2(2),** 126–129. https://doi.org/10.7763/IJFCC.2013.V2.135

Halder A, Misra S. 2014. Semi-supervised fuzzy k-nn for cancer classification from microarray gene expression data. In: 1st International Conference on Automation, Control, Energy and Systems (ACES 2014) (IEEE Computer Society Press). https://doi.org/10.1109/ACES.2014.6808013

Halder A, Kumar A. 2019. Active learning using rough fuzzy classifier for cancer predication from microarray gene expression data. Journal of Biomedical Informatics **92,** p 103136. https://doi.org/10.1016/j.jbi.2019.103136

Huang G, Zhu Q, Siew C. 2006. Extreme learning machine: Theory and applications. Neurocomputing **70(1),** 489–501. https://doi.org/10.1016/j.neucom.2005.12.126

Huang G, Huang GB, Song S, You K. 2015. Trends in extreme learning machines: a review. Neural Networks **61,** 32–48. https://doi.org/10.1016/j.neunet.2014.10.001

Hwang HW, Mendell JT. 2006. Micrornas in cell proliferation, cell death, and tumorigenesis. British journal of cancer **96(6),** 776–780. https://doi.org/10.1038/sj.bjc.6603023

Kumar A, Halder A. 2019. Active learning using fuzzy-rough nearest neighbour classifier for cancer prediction from microarray gene expression data. International Journal of Pattern Recognition and Artificial Intelligence **34(1),** p. 2057001. https://doi.org/10.1142/S0218001420570013

Kumar A, Halder A. 2020. Ensemble-based active learning using fuzzy-rough approach for cancer sample classification. Engineering Applications of Artificial Intelligence **91,** p. 103591. https://doi.org/10.1016/j.engappai.2020.103591

Marak DCB, Halder A, Kumar A. 2021. Semi-supervised ensemble learning for efficient cancer sample classification from miRNA gene expression data. New Generation Computing **39,** 487–513. https://doi.org/10.1007/s00354-021-00123-5

Pirooznia M, Yang J, Yang MQ, Deng Y. 2008. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics **9(1),** 1–13. https://doi.org/10.1186/1471-2164-9-S1-S13

Sung H, Ferlay J, Siegel R L, Laversanne M, Soerjomataram I. 2021. A. Jemal and F. Bray, Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians **71(3),** 209–249. https://doi.org/10.3322/caac.21660

Tarek S, El-Khoribi R, Shoman M. 2017. Gene expression-based cancer classification. Egyptian Informatics Journal **18(3),** 151–159. https://doi.org/10.1016/j.eij.2016.12.001

Vanitha CDA, Devaraj D, Venkatesulu M. 2015. Gene expression data classification using support vector machine and mutual information-based gene selection. Procedia Computer Science **47,** 13–21. https://doi.org/10.1016/j.procs.2015.03.178