Breast cancer diagnosis based on genomic data and extreme learning machine

Nilloofar Jazayeri¹ · Hedieh Sajedi¹

Received: 9 June 2019 / Accepted: 26 November 2019 / Published online: 2 December 2019
© Springer Nature Switzerland AG 2019

Abstract

According to cancer.org news in 2018, the most common cancer diagnosed in women is breast, lung, and colorectal cancers. About 30% of all new cancer diagnoses in women refer to breast cancer. Therefore, predicting breast cancer in its early stages stays a controversial challenge. In this study, we focus on DNA methylation gene expression profiles of patients (series GSE32393 NCBI dataset). For dimension reduction, Non-negative matrix factorization (NMF) is employed and combined with a new method called column-splitting. The reason of superiority of NMF algorithm over other popular ones is its usage for both supervised and unsupervised feature matrix transformation into lower dimensional feature matrix. Afterward the main algorithms that are used for classification are extreme learning machine and support vector machine algorithms. The achieved prediction performances are comparable to deep learning models. The best-attained performance has zero error rate on NCBI 137 samples. While it has shown that, the best deep model on the mentioned data has 2.7 error rate.

Keywords Extreme learning machine · Support vector machine · Dimension reduction · NMF · DNA methylation

1 Introduction

Breast cancer is the most common death cause disease between women after lung cancer. Breast cancer is the most common death cause disease between women after lung cancer. According to Islamic Republic News Agency (IRNA), every year 8000 Iranian people suffer from breast cancer it is around 30–35 persons in every 100,000 people. In addition, about 5–10% of cancer cases derived from genetic causes mostly in BRCA1 and BRCA2 heritable genes. Early diagnosis is difficult and vital. Because the soon they found the signs of the disease, the closer they are to survival from uncontrolled cancer. More than 90% of lumps are not cancer or are benignant ones. In general, doctors do some medical tests for cancer diagnosis [1, 2]. First, they need a complete review of health history. Because some mutations in influential genes in breast

cancer are inherited. To find out if any family members had those mutated genes, physical examination is essential for any visible signs like thickness and softness and the possibility of existing any lumps. After doing those instructions, there are some questionable cases that doctors doubt if they have breast cancer or not and if the answer is yes and which is the level of it. Accordingly, doctors get biopsy tests (In this test they examine a small amount of removed breast tissue under a microscope). There are also other laboratory tests like blood, urine and other genetic tests and imaging tests (X-ray, PET/CT, MRI, ultrasound, etc.). Since those data are using for grouping the samples into low risk and high risk or into multi-stage, the classification algorithms come in handy.

According to doctor's diagnosis data, there are various types of data and in some cases. Some of them may seem unrelated. The most difficult type of patient's data

✉ Hedieh Sajedi, hhsajedi@ut.ac.ir; Nilloofar Jazayeri, N.jazayeri@ut.ac.ir | ¹School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran.

to interpret is the correlation between genes and getting cancer mainly because of the complex relation network between human thousand genes. However, there are some exact genes like BRCA1, BRCA2, and Abraxas. It was observed that mutations in those mentioned genes lead to breast cancer disease.

The best example of effective use of computers in helping doctors with a cancer diagnosis is predicting cancer stages based on gene expression in body cells [3]. A large number of genes and complex relationship between genes and disease, which is not fully discovered yet is another challenge besides the accuracy of prediction. Without machine learning algorithms, diagnosis based on gene expression will be a study of hand selecting genes by experts. To overcome that problem which referred to the curse of dimensionality, dimension reduction techniques are employed.

As we mentioned before, the most motivating reason for using an automated system is obtaining high accuracy in diagnosis, reducing false positive detections and preventing unnecessary cancer surgeries comes as results. A study in the MIT laboratory is an example of the specialist effort that shows how being up to date is essential for data interpretation and correct prediction. Those teams create an algorithm that distinguishes between lymphoma in real time using thousands of pathology data for training the computational model and it is showed to be promising after implementation test results [2].

In this paper, a splitting method is proposed to build some simple classifiers instead of a single complex classifier. Furthermore, we employ dimension reduction methods known as non-negative matrix factorization (NMF). Afterward two classifiers, ELM and SVM, applied to classify data to healthy and cancerous classes and the experiments on the dataset (series GSE32393 NCB1 Dataset) show the superiority of the proposed method compared to the other previous methods.

The rest of the paper is organized as follow: we introduced other works done on diagnosing cancer on different high dimensional datasets in Sect. 2. The proposed method is introduced in Sect. 3. The experimental results are presented in Sect. 4, in which we describe the DNA methylation data sets and illustrate the DNA data analysis results. Finally, we draw some conclusions in Sect. 5.

2 Related works

Machine learning usages in predicting cancer has shown considerable growth in these last two decades. A summary of related works is provided in Table 1. Most of the datasets contain more than thousands features. Other problems are feature redundancy, having noises and

irrelevant data, which cause many troubles in prediction. Studies in [4–7] took advantage from ELM features.

The main limitation in these studies (gene expression classification and clustering subjects) is the curse of dimensionality. NMF method is used in [15] for dimension reduction. Most of the datasets contain more than thousands features. Other problems are feature redundancy, having noises and irrelevant data, which cause many troubles in prediction. Moreover, classification accuracy remains competitive. The study in [16] embeds cost-sensitive factors into classification to overcome unbalance datasets issues. That is important because most of gene expression datasets are unbalance.

In research of [17], a dimension reduction method based on deep neural networks (DNN) was proposed. The DNN is built by four stacked binary restricted Boltzmann machines (RBM). The binary input and output units of the RBM fits the DNA methylation data's bounded support property and the self-learning ability can extract the low dimensional features efficiently and automatically. The experimental results demonstrate the low dimensional features obtained from the proposed DNN can separate the normal samples from the cancer samples effectively with 2.7% error rate. Compared with some recently proposed probabilistic mixture model-based methods, the DNN-based method shows significant advantages.

In [8], Bayesian optimized deep survival models compared with other state of the art machine learning methods for survival like Cox elastic net and random survival forests. The prognostic accuracy of these methods was evaluated in different diseases/datasets (GBMLGG, BRCA, KIPAN) using a high-dimensional transcriptional feature set and a lower-dimensional integrated feature set that combines clinical, genetic, and protein expression features.

There are different data mining techniques that can be used for the prediction of breast cancer recurrence [18]. Researchers analyzed breast cancer data using three classification techniques to predict the recurrence of the cancer and then compared the results [13]. The results indicated that SVM is the best classifier predictor with the test dataset, followed by Artificial Neural Network and Decision Tree. Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

ELM over performed other classification algorithms in some applications [5]. In [5] authors presented a method for classification of three types of brain tumors (i.e., meningioma, glioma, and pituitary tumor) to do so Convolutional Neural Network (CNN) with four convolution layers, four pooling layers and one fully connected layer was used for feature extraction. Then Kernel based ELM was used for classify these features and CNN-KELM had promising

Table 1 Summary of works on predicting cancer

References	Goal	Classifier	Size of dataset	Number of features	Feature extraction method
[8]	Predicting cancer with deep survival models is better than Cox elastic net and RF	Deep learning and Bayesian optimization method	TCGA, BRCA	400–17,000	Elastic net for Cox models (unbiased) input level in DNN
[9]	Lung cancer prediction with deep learning	ConvNet deep learning	2295 MILD dataset (developed method tested on CIFAR10 0000 Image)	256 value 32×32 px image (CIFAR10)	Unsupervised learning [2] and linear SVM classifier
[10]	Predicting lung cancer survivability using ensemble learning	Base learner, bagging, dagging, adaboost, multiboostring, random and SUBSPACE	643,924 samples SEER dataset	149	Correlation-based feature selection [9]
[11]	Cancer prediction using gene expression	C 4.5 decision tree, bagged and boosted decision tree	622 sample of 7 cancer	Differs for each cancer, 7129–24,481 genes	17% Fayyad and Irani's (1993) discretization [12]
[5]	Brain tumor classification via CNN and ELM is better than XGboost, SVM, MLP and KNN	Kernel ELM		233 samples [2] 512×512 px Image	CNN
[13]	Three machine learning techniques for predicting breast cancer recurrence, SVM has least error	C 4.5, SVM and ANN	1189 sample from IBCB	24 clinical feature	none
[14]	Deep learning-based multi-omics integration robustly predicts survival in liver cancer	Deep learning	Whole TCGA data for training the SVM and 360 selected data for deep model	3 omic data mRNA- miRNA	Stacked convolutional auto-encoders

results compared with different classifiers such as SVM, Radial Base Function and some other classifiers.

3 Methodology

Finding an appropriate representation of the data is a challenge issue in many data-analysis tasks. Translating the vast data generated by genomic platforms into accurate predictions of clinical outcomes is a fundamental challenge in genomic medicine. High-dimensional profiles limitation in learning and contemporary platforms like sequencing can provide thousands to millions of features. This leads to the process that experts hand-select a small number of features for training. Prediction models leads to prone to bias and limited by imperfect understanding of disease biology.

Column-splitting method is a preprocessing method, proposed in this paper in order to overcome the curse of dimensionality problem.

In column-splitting we separate the columns of the original data matrix in 28 matrixes with 137×1000 sizes. 27 matrixes transferred to a 137×100 matrix and the size of the last one is 137×578 . Our goal is to reduce the dimension of each matrix from 1000 to 100 (any number less than 137 is acceptable). Since, using NMF for feature selection lead us to a new matrix of data to that the new dimensions is less or equal to the original matrix numbers of rows and column, so we cannot directly reduce the 27,578 feature to a number smaller than 137 because the number of samples are 137 and the transferred space could be the minimum of $\{137, 27,578\}$. Finally, 28 obtained matrixes are combined and sent to classifiers. Figure 1 illustrates the method for split the data column-wised to 28 partitions.

Breast tissue samples were drawn from 114 breast cancers and 23 non-neoplastic breast tissues. The breast cancer tissue samples were from women (mean age 59.4) who were diagnosed with breast cancer. Among the cancers, 33 were at stage 1 and 81 at stage 2/3/4; 34, 59, and 19 were grade 1, 2, and 3 respectively; and 91 were invasive ductal carcinoma breast cancers. The 23 non-neoplastic samples are from healthy woman (mean age 47.6) [20]. DNA methylation profiles across approximately 27,000 CpGs in breast tissues were obtained from women with and without breast cancer. DNA methylation is an epigenetic mechanism that occurs by the addition of a methyl (CH₃) group to DNA, thereby often modifying the function of the genes and affecting gene expression [21]. Epigenetics is the study of heritable changes in gene activity or function that is not associated with any change of the DNA sequence itself [22]. Many researches demonstrated that the DNA methylation, which occurs in the context of a CpG, has strong correlation with diseases, including cancer. There is a strong interest in analyzing the DNA methylation data to find how to distinguish different subtypes of the tumor [17].

The Developed method tested on DNA methylation gene expression profiles came from microarrays chips. Briefly speaking, Methylated DNA happens when a methyl group adds to a nucleotide. The methylated genes can change the risk of getting cancer. An Illustration of Cytosine nucleotide methylation is shown in Figs. 2 and 3 show a DNA methylation profiled representation for a person gene expression obtained from microarray.

The original data set contains 27,578-dimensional data and in some researches like [21] 5000-dimensional data NCBI Data is used after dimension reduction.

The used dataset is on invasive ductal carcinoma breast cancers from NCBI gene expression repository (series GSE32393) with 137 samples and 27,578 feature.

Fig. 1 An illustration of cytosine nucleotide methylation [19]

Used data

Fig. 2 **a** This is the original DNA methylation NCBI series GSE32393 data obtained by microarray. In **(b)** the data splits to 28 separated parts, preparing for dimension reduction, in **(c)** each part trans-

ferred to second space by dimension reduction methods (NMF) and finally in **(d)** the parts assembled in a unit matrix and expressed as the final input data for the classifiers SVM and ELM

3.1 NMF algorithm

We employ a linear map called NMF to transform feature matrix data into another space with fewer dimension. In the following, we summarize the functionality of NMF. In NMF Algorithm a non-negative matrix W is given; the goal is finding nonnegative matrix factors P and Q

$$W_{n \times m} \approx P_{n \times k} Q_{k \times m} \quad (1)$$

The NMF algorithm has an iterative procedure, which, looks for two matrices P and Q to be equal to the original data matrix when they are multiplied together as in Eq. (1). P is feature matrix and Q is weight matrix. Firstly, P and Q are initialized randomly then the mean square error of norm $PQ - W$ is calculated. The conjugate gradient decent method can be used to move in the direction that reaches to the minimum mean square error. The error function is provided in Eq. (2).

$$||W - PQ||^2 = \sum_{ij} W_{ij} - (PQ)_{ij} \quad (2)$$

As it is clear from the equation, the cost function is only convex respect to variable P or only to Q not both. Thus, there is no such algorithm to Minimize $||W - PQ||^2$ with respect to W and H , subject to the constraints $W, H \geq 0$. However used algorithm can find local minima effectively for more detail and the proof of convergence see [24, 25].

Fig. 3 A DNA methylation profiled representation for a person gene expression obtained from microarray. For more detail about Microarray Images, refer to [23]

It is faster than gradient decent around the local minima however, it has more complicated implementation. In study [15] NMF is used for feature extraction.

3.2 Column Splitting method

Column-Splitting method is a preprocessing method, proposed in this paper in order to overcome the curse of dimensionality problem. The following procedure explains its functionality in more detail.

Pseudo code of column-splitting

1. Prepare the numeric data table from DNA methylation microarray's data. Save it into $DATA_{m \times n}$ matrix.
2. Segment the columns of the table into n tables in which all have m rows.
DATA divides into n separate parts with no intersection.

$$|n_1| = |n_2| = \dots = |n_k|, \quad n_i > \text{rank}(DATA), \quad DATA = \bigcup_1^n A^i_{m \times n_k}$$

3. For $i = 1 \rightarrow k$
Obtain B^i by reducing the dimension of each A^i separately using Non-negative Matrix Factorization method. $A^i_{m \times n_j} \rightarrow B^i_{m \times n'_j}, \quad n'_j < n_j$
4. For $i = 1 \rightarrow k$
Concatenate all the B^i 's. $DATA_{new \times p} = [B_1, B_2, \dots, B_k]$

The disadvantages of NMF algorithm occurs when we wanted to directly transfer the DATA into lower dimension space, we could not be able to reach any dimension we want because the Non-negative Matrix Factorization

method always transfer data into second dimension which is less than or equal to the rank of original data.

4 Experimental results

Tables 2 and 3 show the results of column-splitting approach. It should be noted that the error rate in DNN model is 2.7% [17]. Table 2 is devoted to ELM experiments and column-splitting approach. Table 3 is same as previous experiments by NMF but here the classifier is SVM with two different kernel function for nonlinear dataset, which is one of the well-known classification algorithms [26]. We should mention that other SVM parameters tuned automatically by default optimizer used in Matlab implementation (fitcsvm).

In addition, we do not need to set and tune many parameters, which are used in deep model. Because Matlab implementation of each learning model was used in this study and they implementation consider

the optimization process for hyper parameters like kernel function, margin of SVM, and other SVM parameters. However, we used grid search to optimize ELM parameters, the number of hidden nodes and kernel function at last, Kfold cross-validation is used to validate the accuracy of each learning model. After analysis on the results, we can infer that:

1. The process time (Train time and cross fold-validation time) reduced as we reduced the hidden node in hidden layer in ELM.
2. The process time (Train time and cross fold-validation time) reduced as we increased the number of splitting. That is not surprising because as we increased the number of splitting the final data has fewer features.
3. Accuracy rate reduced as we reduced the hidden nodes in the hidden layer in ELM (more than 1000 nodes were good but 500 nodes were not enough to classify data accurately).

Table 2 Dimension reduction and Classification by ELM classifier results on NCBI dataset

Dimension of data	Classifier	Hidden nodes	Split number	Accuracy (%)	Error rate (%)	Time (s)
27,000	ELM	1000	0	100	0	57.3110
27,000	ELM	500	0	99.28	0.72	45.1530
NMF dimension reduction	Sigmoid activation function					
5400	ELM	1000	5	100	0	22.3560
5400	ELM	500	5	99.23	0	13.6690
2700	ELM	1000	10	100	0	14.6450
2700	ELM	500	10	100	0	12.420
900	ELM	1000	30	100	0	14.3360
900	ELM	500	30	100	0	10.0060
540	ELM	1000	50	98.57	1.43	19.2670
540	ELM	500	50	97.85	2.15	14.1210

Table 3 Dimension reduction and classification results on NCBI dataset by SVM classifier

Dimension of data	Classifier	Split number	Accuracy (%)	Error rate (%)	Train time (s)
27,000	SVM	0	84	16	88.0450
NMF dimension reduction	RBF kernel function				
5400	SVM	5	82.6	17.4	68.0450
2700	SVM	10	88.82	11.18	68.0770
900	SVM	30	88.37	11.63	43.5260
540	SVM	50	82.37	17.63	23.3730
NMF dimension reduction	Gaussian kernel function				
5400	SVM	5	98	2	68.0450
2700	SVM	10	97.4	2.6	68.0770
900	SVM	30	99	1	43.5260
540	SVM	50	97	3	23.3730

4. Accuracy rate reduced as we increased the number of pieces of data or the split number (two last experiments for 50 split number).

5 Conclusion

The DNA methylation data can be used to distinguish cancer gene from normal gene. However, the high dimensionality makes it hard to be directly analyzed. Meanwhile, the non-Gaussian properties of the data make many conventional dimension reduction methods do not work well for the clustering task. In this paper, we adopted a dimension reduction method based on NMF algorithm. The experimental result showed that it had promising error rate compare with deep neural networks (DNN).

Compliance with ethical standards

Conflict of interest There is no conflict of interest.

References

1. Chang JC, Wooten EC, Tsimelzon A, Hilsenbeck SG, Gutierrez MC, Elledge RE et al (2003) Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* 362(9381):362–369
2. Conner-Simons A (2015) How a computer can help your doctor better diagnose cancer. CSAIL MIT's Computer Science and Artificial Intelligence Laboratory
3. Salem H, Attiya G, El-Fishawy N (2017) Early diagnosis of breast cancer by gene expression profiles. *Pattern Anal Appl* 20(2):567–578
4. Kumar CA, Ramakrishnan S (2014) Binary classification of cancer microarray gene expression data using extreme learning machines. In: 2014 IEEE international conference on computational intelligence and computing research (ICCIC). IEEE, pp 1–4
5. Pashaei A, Sajedi H, Jazayeri N (2018) Brain tumor classification via convolutional neural network and extreme learning machines. In: 8th International conference on computer and knowledge engineering, Ferdowsi University of Mashhad
6. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: International joint conferences on artificial intelligence, pp 1022–1029
7. Wang K, Duan X, Gao F, Wang W, Liu L et al (2018) Dissecting cancer heterogeneity based on dimension reduction of transcriptomic profiles using extreme learning machines. *PLOS ONE* 13(10):e0205548
8. Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C et al (2017) Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci Rep* 7:11707
9. Ciompi F, Chung K, van Riel SJ, Setio AAA, Gerke PK, Jacobs C et al (2017) Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci Rep* 7:46479
10. Safiyari A, Javidan R (2017) Predicting lung cancer survivability using ensemble learning methods. In: Intelligent systems conference (IntelliSys), London, pp 7–8
11. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 21(20):3896–3904
12. Hall MA (1999) Correlation-based feature selection for machine learning. Ph.D. Thesis, University of Waikato
13. Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR (2013) Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform* 4:124. <https://doi.org/10.4172/2157-7420.1000124>
14. Chaudhary K, Poirion OB, Lu L, Garmire LX (2017) Deep learning-based multi-omics integration robustly predicts survival in liver cancer. <https://doi.org/10.1158/1078-0432.ccr-17-0853>
15. Yuvaraj N, Vivekanandan P (2013) An efficient SVM based tumor classification with symmetry non-negative matrix factorization using gene expression data. In: Information communication and embedded systems (ICICES), International conference on computational intelligence and computing research (ICCIC). IEEE, pp 761–768
16. Liu Y, Lu H, Yan K, Xia H, An C (2016) Applying cost-sensitive extreme learning machine and dissimilarity integration to gene expression data classification. *Comput Intell Neurosc* 2016:8056253
17. Si Z, Yu H, Ma Z (2016) Learning deep features for DNA methylation data analysis. *IEEE Access* 4:2732–2737
18. Jazayeri N, Sajedi H (2019) Early diagnosis of breast cancer based on genomic data using extreme learning machine. In: The international conference on contemporary issues in data science (CIDAS)
19. Nevin C, Carroll M (2015) Sperm DNA methylation, infertility and transgenerational epigenetics. *HSOA J Hum Genet Clin Embryol* 1:004
20. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse32393>. Accessed 24 Sept 2018
21. <https://www.whatisepigenetics.com/dna-methylation>. Accessed 24 September 2018
22. Moore LD, Le T, Fan G (2013) DNA methylation and its basic function. *Neuropsychopharmacology* 38(1):23–38. <https://doi.org/10.1038/npp.2012.112>
23. Li Y, Zhang Y, Li S, Lu J, Chen J, Wang Y et al (2015) Genome-wide DNA methylome analysis reveals epigenetically dysregulated non-coding RNAs in human breast cancer. *Sci Rep* 5:8790
24. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791
25. Yang Z, Hu Y, Liang N, Lv J (2019) Nonnegative matrix factorization with fixed L2-norm constraint. *Circuits Syst Signal Process* 38:3211–3226
26. Bishop CM (2006) Pattern recognition and machine learning. Information science and statistics. Springer, Berlin

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.