International Journal of Engineering

Journal Homepage: www.ije.ir

A Multiple Kernel Learning Based Model with Clustered Features for Cancer Stage Detection using Gene Datasets

A. Mohammadjani, F. Zamani*

Department of Electrical and Computer Engineering, Nooshirvani University of Technology, Babol, Iran

PAPER INFO

Paper history:
Received 06 June 2023
Received in revised form 06 Aguest 2023
Accepted 07 Aguest 2023

Keywords:
Machine Learning
Multiple Kernel Learning
Bioinformatics
Cancer Stage
Dimension Reduction
The Cancer Genome Atlas(TCGA)

ABSTRACT

Genomic data is used in various fields of medicine including diagnosis, prediction, and treatment of diseases. Stage detection of cancer progression is crucial for treating patients because the mortality rate of cancer is higher when it is diagnosed in the late stages. Furthermore, the type of treatment varies depending on the cancer stage. This paper presents a Multiple Kernel Learning based algorithm to predict the stage of cancer using genomic data. Because of the high dimension of genomic data, the curse of dimensionality may degrade the stage prediction. To reduce the dimension, features are clustered first in the proposed algorithm. Then, the original data samples are clustered into smaller subsets with reduced dimensions based on the computed feature clusters. Afterward, for each subset, a kernel matrix is calculated. The kernel matrices are weighted and then combined linearly. Finally, a cancer stage prediction model is trained using the combined kernel matrix and Support Vector Machine. The proposed algorithm is compared with the baseline methods. The classification accuracy of the proposed method outperforms the other methods in 13 cancer groups of 15 from the cancer genome atlas program (TCGA) dataset.

doi: *10.5829/ije.2023.36.11b.08*

1. INTRODUCTION

The increasing generation of genomic data and the need to store, retrieve, and properly analyze them led to the emergence of bioinformatics. Bioinformatics deals with mathematical and computational aspects to understand and process biological data. In other words, the aim of bioinformatics is to increase understanding of biological processes through the use of computational techniques [1].

With the significant growth of biological data generation, they play important role in analyzing and resolving problems in medicine such as cancer diagnosis and treatment [2]. Before the advent of machine learning methods, bioinformatics algorithms were written manually, which made them difficult to be used in applications such as protein structure prediction [3]. Today, machine learning tools and methods are widely used in bioinformatics applications [4].

This paper proposes a machine learning based algorithm to predict the stage of cancer using genomic data. Diagnosing the stage of cancer progression is critical because the mortality rate of cancer is high in its late stages. Furthermore, the type of treatment is different at different stages.

Different types of genomic data are available. The genomic data used in this paper is gene expression, which is commonly used in bioinformatics applications such as cancer diagnosis, treatment, survival, and stage detection.

In this paper, the problem of detecting the stage of cancer progress is considered as a classification problem. The results of machine learning algorithms such as support vector machine, Random Forest, and Multiple Kernel Learning in genomic data classification problems are satisfying. In this paper, we proposed a Feature Clustering Multiple Kernel Learning (FCMKL) algorithm to detect the cancer stage of patients.

*Corresponding Author Institutional Email: zamani@nit.ac.ir (F. Zamani)

Please cite this article as: A. Mohammadjani, F. Zamani, A Multiple Kernel Learning Based Model with Clustered Features for Cancer Stage Detection using Gene Datasets, ***International Journal of Engineering, Transactions B: Applications***, Vol. 36, No. 11, (2023), 2028-2037

The dimension of gene expression data is very high, such that may influence the performance of classification algorithms due to the curse of dimensionality. Curse of dimensionality problem is addressed by dimension reduction. Principle Component Analysis is a widely used method to reduce the dimension of data.

To reduce the gene expression dimension, a novel method is employed in this paper. To this end, the features are clustered first, then data is divided into groups such that in each group, data is represented by the corresponding feature cluster. It is worth noting that the number of clusters and the number of data groups is equal. Finally, in their new representation, data are combined with a Multiple Kernel Learning classifier in order to determine the stage of cancer progression.

The key contributions of the proposed algorithm, Feature Clustering Multiple Kernel Learning (FCMKL), are as follows:

- The genomic data using for cancer stage detection, which is the main focus of this paper, is gene expression. The dimension of gene expression data is high. To avoid the curse of the dimensionality problem, the features are clustered into smaller groups. By grouping features, the classifier does not suffer from the curse of the dimensionality problem because of the reduced dimension of data. Also, this method does not change or remove features.
- For each data group, a kernel matrix is calculated. Then a weighted linear combination of kernel matrices is computed in a Multiple Kernel Learning framework which is used to detect the cancer stage of the patient.
- This paper combines clustering and classification algorithms together to predict the cancer stage of patients.

A block diagram of the proposed method is depicted in Figure 1.

This paper is organized as follows. In the second section, related works are reviewed. The third section explains the proposed algorithm in detail. In section four, the experiment results of the proposed algorithm are demonstrated and discussed. The last section concludes the paper.

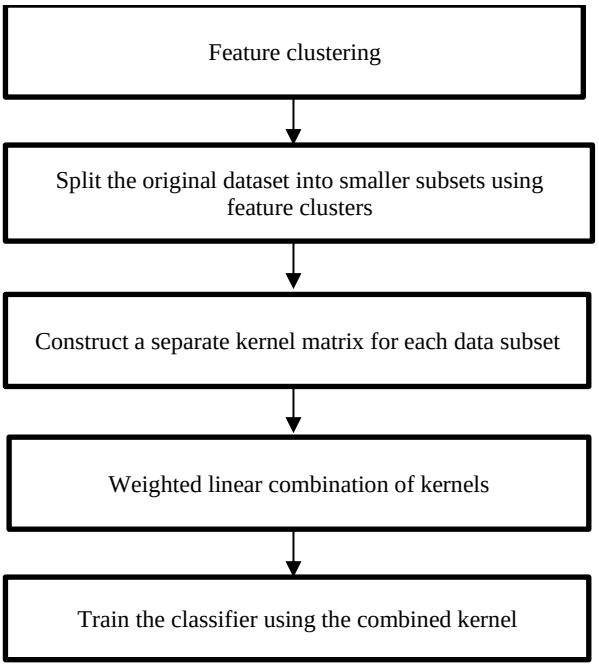

Figure 1. The block diagram of the proposed method

2. RELATED WORKS

This section reviews some works related to machine learning based cancer diagnosis and treatment including cancer stage detection.

An integrated model based on logistic regression and support vector machine for the classification of Colorectal Cancer (CRC) into cancerous and normal samples was proposed by Zhao er al. [5].

The method proposed by Bhalla et al. [6]identifies genes to detect the progress of renal cell cancer. For this purpose, gene expression data from the KIRC cancer group of TCGA dataset is used. This method is based on the fact that there are only a few genes that are important to determine the stage of cancer. In this method, a threshold value is selected for each gene such that determines whether the desired sample is in an early stage or a late one according to the expression of that gene. Finally, the selected genes were fed to the support vector machine.

Huo et al. [7] used the gene expression data for tumor classification based on the sparsity characteristics of genes. To this end, related genes are selected via the sparse group lasso method. Then, tumors are classified by a support vector machine. Ranjani Rani and Ramyachitra [8] proposed a similar framework for cancer classification. They employed the Spider Monkey Optimization algorithm to select related genes.

After detecting differentially expressed genes, in the framework proposed by Xu etal. [9], a Protein Preotein Interaction (PPI) network based neighborhood scoring technique was used in combination with Support Vector Machine for colon cancer diagnosis and recurrence prediction.

Medjahed et al. [10] employed Support Vector Machine in two phases to select the best gene set of DNA microarray for cancer diagnosis task. A two-stage feature selection method based on Multiple Kernel Learning method was proposed by Du et al. [11] to predict cancer. In the first step of the proposed method, relevant features are identified by Multiple Kernel Learning. In the second step, a subset of features from the set of candidate features obtained in the first step is specified.

Data fusion based on Multiple Kernel Learning is proposed by Speicher and Pfeifer [12] to identify cancer subtypes. In order to reduce the dimension of gene data, the proposed method was combined with a graph embedding framework.

A model based on the combination of clustering and Multiple Kernel Learning framework was proposed by Speicher and Pfeifer [13] to identify cancer subtypes. In the proposed model, the features are clustered based on the combination of several kernels, then the effect of each feature cluster on a patient cluster is measured.

The proposed method by Tao et al. [14] deals with the classification of five subtypes of breast cancer based on Multiple Kernel Learning. The data used in this research are gene expression, DNA methylation, and copy number variation from the TCGA dataset. Some genes may have little or no effect on the classification of breast cancer subtypes, which should be identified. For this purpose, the p-values of genes were calculated using the Wilcoxon rank sum. The Benjamini-Hochberg false discovery rate is then determined to adjust the computed p-values. Genes with p-value less than 0.05 are selected as significant genes.

Four types of genomic data in addition to pathological images were used by Sun et al. [15] to predict the survival of breast cancer patients. In the proposed method, Multiple Kernel Learning is employed to integrate different data types.

To predict the survival of patients with squamous cell lung cancer who underwent surgery, a new method based on Multiple Kernel Learning is proposed by Zhang et al. [16]. Due to the small number of samples, to deal with the problem of the curse of dimensionality, a linear correlation algorithm is employed to select the optimal features.

Multiple Kernel Learning was used by Wilson et al. [17] to determine the best kernels calculated from two types of data, including clinical data and microRNA from the TCGA dataset. The goal is to predict whether a patient with ovarian cancer would live more than three years after diagnosis or not.

A method to determine the cancer stage using Multiple Kernel Learning was proposed by Rahimi and Gönen [18]. In this paper, instead of identifying clusters of gene expression features and computing kernel matrices, it was proposed to combine these two steps into a single model using prior knowledge about pathways and sets of genes. For this purpose, they create a separate kernel matrix for each gene set, then combine them using a Multiple Kernel Learning algorithm.

A set of pathways/genes along with gene data were used by Rahimi and Gönen [19] to detect the cancer stage. Different types of cancers with distinct biological mechanisms, have similarities. In this paper, each cancer group is considered as a specific task. A multi-task learning formula is used in which different tasks are being trained simultaneously. In fact, the goal is to identify similarities between cancer groups (i.e., tasks) in terms of their basic mechanisms. Joint clustering is used for this purpose.

Deep learning based methods generally have very high accuracy in data classification. Zohrevand et al. [20] introduced Convolutional Neural Network, which is a powerful deep learning approach, that was employed to Finger-Knuckle-Print recognition. A Fully Convolutional Network in combination with the graph's shortest layer path has been used for fluid segmentation in retina images [21]. Also, a fully automated model was trained by Azimi et al. [22] for fluid segmentation. In this two-path method, the first and last layers of the retina are segmented in the Neutrosophic domain. Then, a Fully Convolutional Network is used for fluid segmentation. Assigning appropriate values for parameters is very important in machine learning based methods. Chegeni et al. [23] proposed a mathematical model to compute the Convolutional Neural Network model parameters automatically. Deep learning based methods suffer from high computational complexity in the training phase and a large number of parameters including weights. To address the mentioned problems a compact version of the Convolutional Neural Network which is called SqueezeNet is employed for document classification while its classification results were comparable to Convolutional Neural Network [24].

Salimy et al. [25] proposed a deep learning framework to predict the survival of colon cancer patients. This method integrates three types of genomic data including gene expression, DNA methylation, and clinical data by autoencoder. Slimene et al. [26] used microRNA for cancer classification. After converting the microRNA data into images, ResNet, which is a pretrained Deep Neural Network is employed to classify data.

3. PROPOSED FCMKL

This paper focuses on diagnosing the early and late cancer stages by using a gene expression data set. Cancer stage detection is considered as a binary classification problem. In the problems like cancer stage detection in which data samples are usually not separable, the use of the kernel function, which implicitly maps data to a high dimension space, improves the classification accuracy (Figure 2a).

The dimension of data samples in the gene expression dataset is very high, which degrades the performance of cancer stage classification due to the curse of dimensionality. To address this problem, it is necessary to reduce the data dimension (Figure 2b).

In order to reduce the data dimension, features are clustered in the first step of the proposed method. The

idea is to compute a separate kernel for each cluster of features. After reducing the feature dimension, a Multiple Kernel Learning classifier is trained to classify cancer stages by using the computed kernels (Figure 2c).

Figure 2 illustrates three different ways to compute kernel matrix for high dimension Gene Expression data. (a) In this case, a kernel function is used to compute kernel matrix simply. Since the dimension of Gene Expression Data is high, cure of dimensionality problem will reduce classification accuracy in this method. (b) To address the curse of dimensionality problem, it is recommanded to reduce the dimension of data by employing dimension reduction algorithms like PCA before computing kernel. (c) Another approch to reduce the dimension of Gene Expression Data, is to cluster features. This method, which is used in this paper, does not change or remove features.

Suppose the dataset contains N data samples and the feature dimension of each data sample is d. The features are clustered into c clusters. Therefore, each cluster contains N data samples which are d/c dimensional. For each d/c-dimensional cluster, a separate $N \times N$ kernel is computed.

The ratio of the number of features to the number of samples is determinanat in the classification performance. It should be noted that for a fixed sample size, if the number of features grows, the classification error will decrease first and then will increase [27]. In the case that the features are independent, its enough that the number of features does not exceed N-1. As the feature correlation increases, this number decreases such that if the correlation is very high, this number decreases to $\sqrt{N}$ which is used as the number of clusters in the proposed method [27].

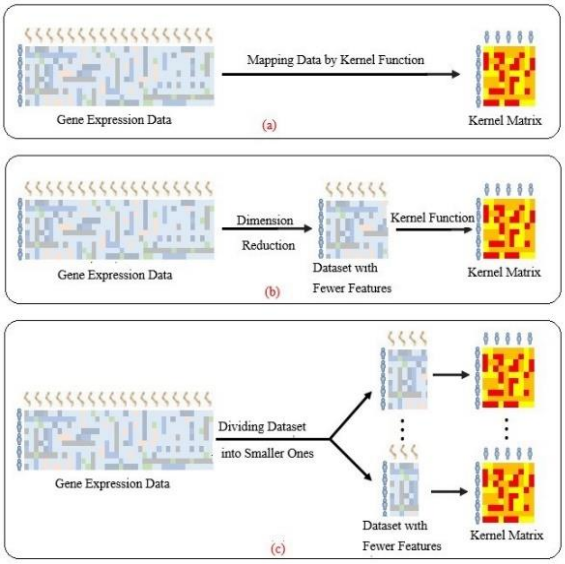

Figure 2. Kernel matrix for high dimension Gene Expression data

The architecture of FCMKL is illustrated in Figure 3. In the following sections, the training and testing phases are explained in detail.

Figure 3 illustrates the architecture of FCMKL. First, the dataset is divided into training and testing sets. In the training phase, the features of the training data are clustered. Then, a kernel matrix is computed for each *feature* cluster. A single kernel is obtained by weighted linear combination of computed kernels. Then, the kernel based Support Vector Machine is trained. In the testing phase, after calculating the kernel corresponding to the testing data based on the feature clusters detected in the training phase, the testing data are classified by the trained support vector machine.

3. 1. Training Phase Following are the main steps of the FCMKL algorithm.

Step 1: Feature clustering

As described before, to address the curse of dimensionality problem in the proposed algorithm, the original data set is divided into smaller ones. To this end, the features are clustered by the k-means clustering algorithm. More precisely, the rows (samples) and columns (features) of the data set are interchanged and given as input to the kmeans algorithm. Kmeans algorithm clusters features based on samples, the output of which is feature clusters.

Step 2: Split the dataset into smaller datasets based on feature clusters

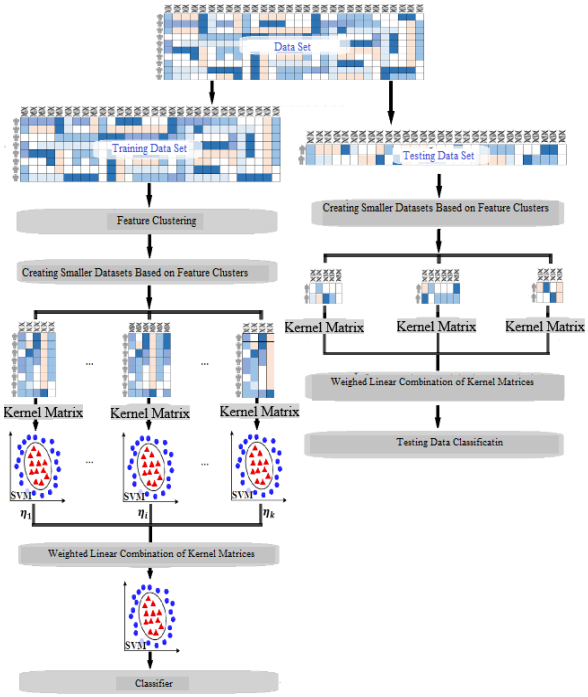

Figure 3. FCMKL architecture

By using the feature clusters obtained in the previous step, the dataset is divided into smaller subsets such that each subset uses one feature cluster. In each small subset, the rows are samples of the original dataset and the columns are the features in the corresponding cluster. In this way, there will be data subsets that have *N* samples wit different features.

Step 3: Computing the kernel matrix for each data subset
Each data subset is implicitly mapped to a high-dimensional feature space by using a separate kernel function. Eventually, for each data subset, an N×N kernel matrix is computed based on the corresponding kernel function.

Step 4: Kernel weighting
Before combining kernels calculated in the previous step, we should weight them. To this end, the AUC classification accuracy of each kernel based on Support Vector Machine is computed. Then, weights are assigned to the each kernel by using Equation (1).

$$\eta_i = \frac{(AUC_i)^2}{\sum_{j=1}^{P}(AUC_j)^2} \quad (1)$$

where η_i is the weight of the i^{th} kernel, P is the number of kernels and AUC_i is the result of predicting the cancer stage by using i^{th} kernel.

Step 5: Weighted linear combination of kernel matrices
In this step, the kernels are linearly combined based on the weights calculated in the previous step. Then the kernel matrices are combined according to Equation (2) and a single kernel matrix is created.

$$k_s(x_i, x_j) = \sum_{l=1}^{p} \eta_l k_l(x_i, x_j) \quad (2)$$

where $k_l(x_i, x_j)$ is the l^{th} kernel matrix and $k_s(x_i, x_j)$ is the combined kernel matrix.

Step 6: Training kernel-based support vector machine
Finally, by using the combined kernel matrix, a kernel based Support Vector Machine is trained.

3. 2. Testing Phase The proposed method is evaluated by measuring the testing data classification accuracy. The main steps of the testing phase are as follows:

Step 1: Split the testing dataset into smaller datasets based on feature clusters
In first step, using the feature clusters obtained in the training phase, the testing set is divided into smaller subsets.

Step 2: Constructing training-testing kernel matrices
The kernel matrices of training-testing data are calculated in this step using training and testing subsets.

Step 3: Weighted linear combination of training-testing kernel matrices
The training-testing kernels computed in the previous step are combined using the weights calculated in the training phase by Equation (2).

Step 4: Classification of testing data using kernel-based Support Vector Machine
Finally, by using the combined training-testing kernel matrix, testing data are classified by trained Support Vector Machine.

4. EXPERIMENTAL RESULTS

In this section, some experiments have been conducted to evaluate the performance of the proposed algorithm using the TCGA dataset. Then, the proposed method is compared with some baseline methods.

4. 1. TCGA Dataset In the experiments, several groups of cancers available in the TCGA dataset were used to detect the cancer stage. In this dataset, gene expression values of cancer patients, which includes more than 10,000 tumors, are available. In the experiments, HTSeq-FPKM records including primary tumors have been downloaded and used for each disease group.

The TCGA database includes clinical annotations for cancer patients. One of the annotated items, is the degree of cancer progression, which is a number between 1 and 4 for each patient.

Due to the fact that it is clinically significant to distinguish between early and late stages of cancer, in this paper, primary tumors annotated with stage 1 are considered as early stage and the remaining tumors annotated with stages 2, 3, and 4 are considered as late. Disease group information used in this paper is summarized in Table 1.

4. 2. Experiment Settings For each cancer group, 80% of tumors were selected as training data and the remaining 20% as testing data. The data was divided in such a way that the proportion of positive and negative classes in the training and testing sets is almost equal.

The rane of gene expression value is large. After adding a fixed value, the gene expression values have been converted to a more limited range using log 2. The training set was normalized to have zero mean and standard deviation of one, and then the testing set was as well.

The effiiency of the proposed algorithm is compared with upport Vector Machine, Random Forest, combination of PCA and Support Vector Machine, Deep Neural Network and also Multiple Kernel Learning using Halmark gene dataset which includes 50 gene sets [18]. It was extracted from some molecular databases. Each gene set contains information about a specific biological state or a biological process. Rahimi and Gönen [19] divided the gene expression dataset into 50 smaller ones based on the features available in the Hallmark gene set.

TABLE 1. Summary of 15 cancer groups in the TCGA dataset

Cancer Group	RF	SVM	PCA+SVM	DNN	MKL[H]	FCMKL
BRCA	0.55	0.62	0.64	0.64	0.63	0.65
COAD	0.58	0.65	0.66	0.67	0.68	0.71
ESCA	0.74	0.67	0.69	0.67	0.71	0.81
HNSC	0.53	0.67	0.67	0.67	0.69	0.79
KICH	0.69	0.66	0.69	0.69	0.65	0.86
KIRC	0.76	0.75	0.75	**0.78**	0.75	0.77
KIRP	0.79	0.78	0.79	0.77	0.80	0.81
LIHC	0.67	0.65	0.65	0.65	0.65	0.68
LUAD	0.63	0.62	0.62	0.58	0.62	0.64
LUSC	0.63	0.62	0.62	0.64	0.62	0.65
PAAD	0.67	0.69	0.71	0.68	0.74	0.80
READ	0.64	0.55	0.61	0.53	0.63	0.72
STAD	0.71	0.72	0.70	0.67	0.69	0.76
TGCT	0.73	0.76	0.75	0.68	0.72	0.82
THCA	0.68	0.67	0.68	**0.72**	0.68	0.70
MEAN	0.67	0.67	0.68	0.67	0.68	0.74

To implement random forest, the randomForestSRC package was used [28]. The number of trees for this algorithm was selected from the set {500, 1000, 1500, 2000, 2500} using 4-fold cross validation.

The code shared by Ma et al. [29] was also used to implement the deep neural network.

To implement Support Vector Machine and Multiple Kernel Learning using Hallmark gene set, code shared by Rahimi and Gönen [18] was used and the MOSEK package is used to solve the quadratic optimization problems[1].

To compute kernel matrices, Gaussian kernel function was used:

$$k(x_i, x_j) = \exp\left(-\frac{(x_i-x_j)^T(x_i-x_j)}{2\sigma^2}\right) \quad (3)$$

such that σ, the kernel width parameter, was set to the average euclidean distance between all pairs of training data.

In the proposed algorithm, the regularization parameter C was set to 1. Moreover, as discussed in section 3, since genomic data have a very high correlation [18, 19], the number of clusters should be equal to the ratio of the number of features in the dataset to the square root of the number of training data samples [27]. In this way, the features are clustered in such a way that the average number of features in each feature cluster is equal to the square root of the number of training data samples as proposed by Zhang et al. [16].

To compare classification performance of the mentioned algorithms, the evaluation measurement AUC (area under the ROC curve) has been calculated.

To achieve more reliable results, all the experiments were repeated 100 times, and the average of the AUC values were reported (Table 2).

Also, the results of the experiments are illustrated and compared in Figure 4. As Figure 4 shows, the average performance of the algorithms in all datasets is better than random case (in which AUC equal to 0.5). Therefore, the gene expression dataset has significant information about the stages of cancers.

By comparing the classification accuracy of PCA+SVM algorithm with SVM, it is observed that PCA+SVM achieved better results in 8 of 15 cancer groups, while SVM was better in only two groups. The greatest performance improvement of PCA+SVM was in READ cancer group (6%), and the greatest performance reduction was in STAD cancer group (2%).

By comparing the classification accuracy of FCMKL algorithm with RF, it is observed that FCMKL achieved better results in all 15 cancer groups. Performance improvement was significantly better in all groups. For example, compared to RF, FCMKL has improved the classification performance of BRCA by 10%, HNSC by 26%, KICH by 17%, PAAD and COAD by 13%, READ by 8%, STAD by 5%, and TGCT by 9%.

By comparing the classification accuracy of FCMKL algorithm with SVM, it is observed that FCMKL has

[1] https://www.mosek.com/

TABLE 2. Average of AUC values of Random Forest (RF), Support Vector Machine (SVM), Multiple Kernel Learning using Hallmark dataset (MKL[H]), combination of PCA and SVM (PCA+SVM), Deep Neural Network (DNN) and proposed algorithm (FCMKL) on 15 cancer groups from the TCGA dataset

Cancer Group	Cancer Name	Early Stage	Late Stage	Total
BRCA	Adrenocortical carcinoma	202	995	1197
COAD	Colon adenocarcinoma	85	422	507
ESCA	Esophageal carcinoma	21	130	151
HNSC	Head and neck squamous carcinoma	27	450	477
KICH	Kidney chromophobe	29	60	89
KIRC	Kidney renal clear cell carcinoma	297	311	608
KIRP	Kidney renal papillary cell carcinoma	187	105	292
LIHC	Liver hepatocellular carcinoma	191	201	392
LUAD	Lung adenocarcinoma	324	261	585
LUSC	Lung squamous cell carcinoma	271	276	547
PAAD	Pancreatic adenocarcinoma	21	158	179
READ	Rectum adenocarcinoma	34	132	166
STAD	Stomach adenocarcinoma	59	324	383
TGCT	Testicular germ cell tumors	56	26	82
THCA	Thyroid carcinoma	321	245	566

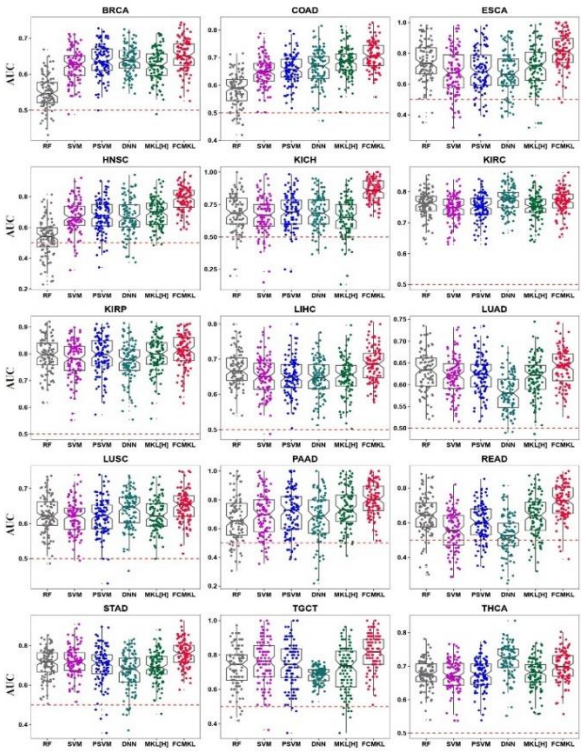

Figure 4. Performance comparison of Random Forest (RF), Support Vector Machine (SVM), Multiple Kernel Learning using Hallmark dataset (MKL[H]), combination of PCA and SVM (PCA+SVM), Deep Neural Network (DNN) and proposed algorithm (FCMKL) on 15 cancer groups from TCGA dataset. The box and dot plots compare the averages AUC values. Orange dashed lines indicate baseline performance level (AUC = 0.5)

obtained better results in all 15 cancer groups. Performance improvements in some groups have been remarkable. Compared to SVM, FCMKL has improved the classification performance of ESCA by 14%, HNSC by 12%, KICH by 20%, PAAD by 11%, READ by 17%, COAD and TGCT by 6%.

By comparing the classification accuracy of FCMKL algorithm with PCA+SVM, it is observed that FCMKL has obtained better results in all 15 groups. Compared to PCA+SVM, FCMKL has improved the classification performance of ESCA and HNSC by 12%, KICH by 17%, PAAD by 9%, READ by 11%, TGCT by 7%, and STAD by 6%.

By comparing the classification accuracy of FCMKL algorithm with DNN, it is observed that FCMKL has obtained better results in 13 of 15 data sets, while DNN was better in only two data sets. The greatest performance improvement of FCMKL was 19%, in READ cancer group and the greatest performance reduction was 2% in THCA cancer group.

By comparing the classification performance of FCMKL algorithm with Multiple Kernel Learning using Hallmark dataset [18], it is observed that FCMKL has obtained better results in all 15 groups. The performance improvement on some datasets has been significantly better. Compared to Rahimi and Gönen's work [18], FCMKL improved the classification performance of HNSC and ESCA by 10%, KICH by 21%, READ by 9% and STAD by 7%, TGCT by 10% and PAAD by 6%.

Table 3 shows the number of clusters computed in the FCMKL algorithm for each cancer group.

Figure 5 shows the number of cluster members in the proposed algorithm for 15 cancer groups. The dot chart and the violin chart display the range, mean and distribution of the number of cluster members. Each black dot represents a cluster. The red dot indicates the average number of cluster members for each cancer group. The largest cluster has 234 members and the smallest one has one member. Considering that the total number of features is more than 19,000, the distribution of features in the clusters seems to be appropriate.

The computed weights for each kernel in the FCMKL algorithm for 15 cancer groups are shown in Figure 6. Violin and box plots represent the range, mean, and weight distribution of kernels. Since the maximum weight assigned to the kernels is equal to 0.0026, we can

TABLE 3. The number of clusters computed in the FCMKL algorit0hm for each cancer group

Cancer Group	Cluster Number
THCA	921
TGCT	2393
STAD	1120
READ	1680
PAAD	1624
LUSC	937
LUAD	906
LIHC	1104
KIRP	1281
KIRC	890
KICH	2286
HNSC	1004
ESCA	1785
COAD	973
BRCA	635

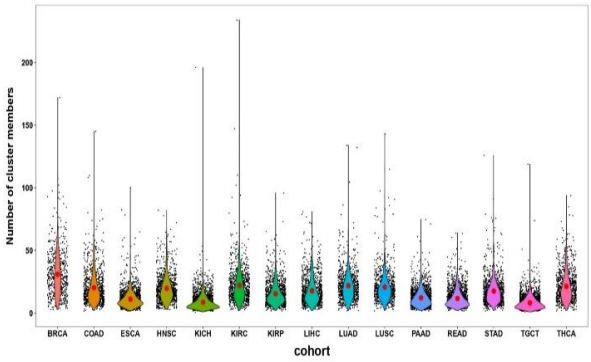

Figure 5. The number of cluster members in the FCMKL algorithm for each cancer group is shown in this figure. The black dots represent the clusters and the red dots represent the average number of cluster members in each group. The violin diagram shows the range and distribution of the clusters in terms of the number of their members

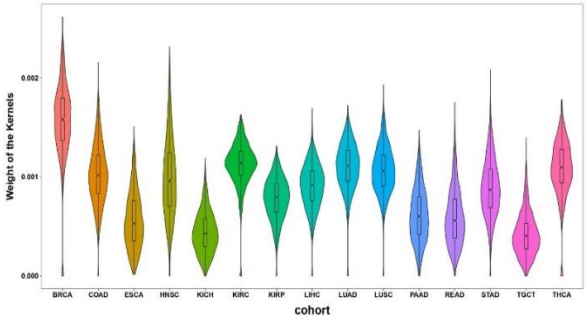

Figure 6. Computed weights for each kernel in the FCMKL algorithm for each cancer group are shown in this figure. Violin diagram and box diagram show the range, mean and weight distribution of kernels

conclude the majority of kernels are effective in classification. The maximum number of zero kernel weights in a cancer group is 20.

The proposed algorithm is implemented by R language. The experiments are conducted on a Windows 10 system, which contains a core i7 CPU with 8 cores and 16GB RAM. The training time of the proposed algorithm varies between 15 minutes to one hour and 10 minutes for different cancer groups.

5. CONCLUSION

Genomic data are useful in many medical applications including disease diagnosis, prevention and treatment. Cancer is one of the most dangerous and life-threatening diseases in the world and is considered as one of the most important causes of death. It is vital to detect the stage of cancer in a patient because if the disease is detected at an early stage, it will be curable. Also, the type of treatment is different in different stages of the disease.

In this paper, an algorithm, FCMKL, is proposed to improve cancer stage detection using feature clustering based Multiple Kernel Learning. Due to the fact that genomic data have a very high dimension, we are facing the problem of the curse of dimensionality. To address this problem, the features of the original dataset are first clustered based on samples. Then, using feature clusters, the original dataset which has a high dimension is divided into smaller datasets in terms of the number of features. For each of these smaller data sets, a kernel matrix is computed. The kernel matrices are weighted and linearly combined. Finally, using the resulting kernel matrix, the Support Vector Machine is trained to determine the cancer stage. The experiments indicate promising performance of the proposed algorithm.

Employing another clustering algorithms may result in reducing the number of clusters. By reducing the number of clusters, the computation time will decrease. Also, there are another genomic data type like microRNA and DNA methylation which we did not used in our proposed method. By using multimodal data, the classification accuracy will increase.

6. REFERENCES

1. Chicco, D., "Ten quick tips for machine learning in computational biology", ***BioData Mining***, Vol. 10, No. 1, (2017), 35. doi: 10.1186/s13040-017-0155-3.
2. Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K. and Zhou, Y., "Sixty-five years of the long march in protein secondary structure prediction: The final stretch?", ***Briefings in Bioinformatics***, Vol. 19, No. 3, (2018), 482-494. doi: 10.1093/bib/bbw129.
3. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J.A., Armananzas, R., Santafé, G. and Pérez, A.,

"Machine learning in bioinformatics", ***Briefings in Bioinformatics***, Vol. 7, No. 1, (2006), 86-112. doi: 10.1093/bib/bbk007.

4. Wang, J.T., Zaki, M.J., Toivonen, H.T. and Shasha, D., Introduction to data mining in bioinformatics, in Data mining in bioinformatics. 2005, Springer.3-8.
5. Zhao, D., Liu, H., Zheng, Y., He, Y., Lu, D. and Lyu, C., "A reliable method for colorectal cancer prediction based on feature selection and support vector machine", ***Medical & Biological Engineering & Computing***, Vol. 57, (2019), 901-912. doi: 10.1007/s11517-018-1930-0.
6. Bhalla, S., Chaudhary, K., Kumar, R., Sehgal, M., Kaur, H., Sharma, S. and Raghava, G.P., "Gene expression-based biomarkers for discriminating early and late stage of clear cell renal cancer", ***Scientific Reports***, Vol. 7, No. 1, (2017), 44997. doi: 10.1038/srep44997.
7. Huo, Y., Xin, L., Kang, C., Wang, M., Ma, Q. and Yu, B., "Sgl-svm: A novel method for tumor classification via support vector machine with sparse group lasso", ***Journal of Theoretical Biology***, Vol. 486, (2020), 110098. doi: j.jtbi.2019.110098.
8. Rani, R.R. and Ramyachitra, D., "Microarray cancer gene feature selection using spider monkey optimization algorithm and cancer classification using svm", ***Procedia Computer Science***, Vol. 143, (2018), 108-116. doi: 10.1016/j.procs.2018.10.358.
9. Xu, G., Zhang, M., Zhu, H. and Xu, J., "A 15-gene signature for prediction of colon cancer recurrence and prognosis based on svm", ***Gene***, Vol. 604, (2017), 33-40. doi: 10.1016/j.gene.2016.12.016.
10. Medjahed, S.A., Saadi, T.A., Benyettou, A. and Ouali, M., "Kernel-based learning and feature selection analysis for cancer diagnosis", ***Applied Soft Computing***, Vol. 51, (2017), 39-48. doi: 10.1016/j.asoc.2016.12.010.
11. Du, W., Cao, Z., Song, T., Li, Y. and Liang, Y., "A feature selection method based on multiple kernel learning with expression profiles of different types", ***BioData mining***, Vol. 10, No. 1, (2017), 1-16. doi: 10.1186/s13040-017-0124-x.
12. Speicher, N.K. and Pfeifer, N., "Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery", ***Bioinformatics***, Vol. 31, No. 12, (2015), i268-i275. doi: 10.1093/bioinformatics/btv244.
13. Speicher, N.K. and Pfeifer, N., "An interpretable multiple kernel learning approach for the discovery of integrative cancer subtypes", arXiv preprint arXiv:1811.08102, (2018). doi: 10.48550/arXiv.1811.08102.
14. Tao, M., Song, T., Du, W., Han, S., Zuo, C., Li, Y., Wang, Y. and Yang, Z., "Classifying breast cancer subtypes using multiple kernel learning based on omics data", ***Genes***, Vol. 10, No. 3, (2019), 200. doi: 10.3390/genes10030200.
15. Sun, D., Li, A., Tang, B. and Wang, M., "Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome", ***Computer Methods and Programs in Biomedicine***, Vol. 161, (2018), 45-53. doi: 10.1016/j.cmpb.2018.04.008.
16. Zhang, A., Li, A., He, J. and Wang, M., "Lscdfs-mkl: A multiple kernel based method for lung squamous cell carcinomas disease-free survival prediction with pathological and genomic data", ***Journal of Biomedical Informatics***, Vol. 94, (2019), 103194. doi: 10.1016/j.jbi.2019.1031.
17. Wilson, C.M., Li, K., Yu, X., Kuan, P.-F. and Wang, X., "Multiple-kernel learning for genomic data mining and prediction", ***BMC bioinformatics***, Vol. 20, (2019), 1-7. doi: 10.1186/s12859-019-2992-1.
18. Rahimi, A. and Gönen, M., "Discriminating early-and late-stage cancers using multiple kernel learning on gene sets", ***Bioinformatics***, Vol. 34, No. 13, (2018), i412-i421. doi: 10.1093/bioinformatics/bty239.
19. Rahimi, A. and Gönen, M., "A multitask multiple kernel learning formulation for discriminating early-and late-stage cancers", ***Bioinformatics***, Vol. 36, No. 12, (2020), 3766-3772. doi: 10.1093/bioinformatics/btaa168.
20. Zohrevand, A., Imani, Z. and Ezoji, M., "Deep convolutional neural network for finger-knuckle-print recognition", ***International Journal of Engineering, Transactions A:Bascs***, Vol. 34, No. 7, (2021), 1684-1693. doi: 10.5829/ije.2021.34.07a.12.
21. Azimi, B., Rashno, A. and Fadaei, S., "Fully convolutional networks for fluid segmentation in retina images", in 2020 International Conference on Machine Vision and Image Processing (MVIP), IEEE. (2020), 1-7.
22. Azimi, B., Rashno, A. and Fadaei, S., "Two-path neutrosophic fully convolutional networks for fluid segmentation in retina images", ***AUT Journal of Modeling and Simulation***, Vol. 54, No. 1, (2022), 85-104. doi: 10.22060/miscj.2022.21258.5277.
23. Chegeni, M.K., Rashno, A. and Fadaei, S., "Convolution-layer parameters optimization in convolutional neural networks", ***Knowledge-Based Systems***, Vol. 261, (2023), 110210. doi: 10.1016/j.knosys.2022.110210.
24. Hassanpour, M. and Malek, H., "Learning document image features with squeezenet convolutional neural network", ***International Journal of Engineering, Transactions A:Bascs***, Vol. 33, No. 7, (2020), 1201-1207. doi: 10.5829/ije.2020.33.07a.05.
25. Salimy, S., Lanjanian, H., Abbasi, K., Salimi, M., Najafi, A., Tapak, L. and Masoudi-Nejad, A., "A deep learning-based framework for predicting survival-associated groups in colon cancer by integrating multi-omics and clinical data", ***Heliyon***, Vol. 9, No. 7, (2023). doi: 10.1016/j.heliyon.2023.e17653.
26. Slimene, I., Messaoudi, I., Oueslati, A.E. and Lachiri, Z., "Deep learning-based cancer disease classification through microrna expression", in 2022 IEEE Information Technologies & Smart Industrial Systems (ITSIS), IEEE. (2022), 1-6.
27. Hua, J., Xiong, Z., Lowey, J., Suh, E. and Dougherty, E.R., "Optimal number of features as a function of sample size for various classification rules", ***Bioinformatics***, Vol. 21, No. 8, (2005), 1509-1515. doi: 10.1093/bioinformatics/bti171.
28. Ishwaran, H. and Kogalur, U.B., "Fast unified random forests for survival, regression, and classification (rf-src)", ***R Package Version***, Vol. 2, No. 1, (2019).
29. Ma, B., Meng, F., Yan, G., Yan, H., Chai, B. and Song, F., "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data", ***Computers in Biology and Medicine***, Vol. 121, (2020), 103761. doi: 10.1016/j.compbiomed.2020.103761.

Persian Abstract

چکیده

داده های ژنومی در زمینه های مختلف پزشکی از جمله تشخیص، پیش بینی و درمان بیماری ها استفاده می شود. تشخیص میزان پیشرفت سرطان برای درمان بیماران بسیار مهم است زیرا میزان مرگ و میر سرطان زمانی که در مراحل پایانی تشخیص داده می شود، بیشتر است. علاوه بر این، نوع درمان با توجه به میزان پیشرفت این بیماری متفاوت است. این مقاله یک الگوریتم مبتنی بر یادگیری چند هسته‌ای برای پیش‌بینی میزان پیشرفت سرطان با استفاده از داده‌های ژنومی پیشنهاد می‌کند. به دلیل ابعاد بالای داده های ژنومی، نفرین ابعاد ممکن است دقت پیش بینی پیشرفت سرطان را کاهش دهد. برای کاهش ابعاد، در الگوریتم پیشنهادی ویژگی‌ها ابتدا خوشه‌بندی می‌شوند. سپس، نمونه‌های داده اصلی به زیر مجموعه‌های کوچک‌تر با ابعاد کاهش‌یافته بر اساس خوشه‌های ویژگی محاسبه‌شده خوشه‌بندی می‌شوند. پس از آن، برای هر زیر مجموعه، یک ماتریس هسته ساخته شده به آنها وزن اختصاص داده می‌شود. سپس ماتریس‌های وزن‌دار به صورت خطی ترکیب می شوند. در نهایت، یک مدل پیش‌بینی میزان پیشرفت سرطان با استفاده از ماتریس هسته ترکیبی و ماشین بردار پشتیبان آموزش داده می‌شود. الگوریتم پیشنهادی با روش های پایه مقایسه شده است. دقت طبقه‌بندی روش پیشنهادی از روش‌های دیگر در ۱۳ گروه سرطانی از ۱۵ گروه مجموعه داده TCGA بهتر است.