Journal of Biomedical Informatics 95 (2019) 103213

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A novel feature selection method for microarray data classification based on hidden Markov model

Mohammadreza Momenzadeh[a], Mohammadreza Sehhati[a,b,*], Hossein Rabbani[a,b]

[a] Department of Bioelectric and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
[b] Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

ARTICLE INFO

Keywords:
Cancer classification
DNA microarray
Feature selection
Hidden Markov model (HMM)
Multi-criteria ranking

ABSTRACT

In this paper, a novel approach is introduced for integrating multiple feature selection criteria by using hidden Markov model (HMM). For this purpose, five feature selection ranking methods including Bhattacharyya distance, entropy, receiver operating characteristic curve, *t*-test, and Wilcoxon are used in the proposed topology of HMM. Here, we presented a strategy for constructing, learning and inferring the HMM for gene selection, which led to higher performance in cancer classification. In this experiment, three publicly available microarray datasets including diffuse large B-cell lymphoma, leukemia cancer and prostate were used for evaluation. Results demonstrated the higher performance of the proposed HMM-based gene selection against Markov chain rank aggregation and using individual feature selection criterion, where applied to general classifiers. In conclusion, the proposed approach is a powerful procedure for combining different feature selection methods, which can be used for more robust classification in real world applications.

1. Introduction

DNA Microarray is a well-known technology that allows the simultaneous observation of the expression levels of thousands of genes. In the recent years, gene expression data is widely used in different fields including medicine and especially cancer. In cancer classification, finding hidden patterns in the expression profiles can increase the prediction accuracy. Meanwhile, curse of dimensionality problem remains as a main challenge. Dimensionality reduction methods are generally divided into two categories: feature extraction or transform-based methods [1–3] and feature selection methods [4–7] which are considered in this study. Feature selection is used for removing redundant and irrelevant features which are not informative and do not improve the classification performance. The benefits of feature selection include: (1) providing a low-complexity model by reducing model parameters, (2) avoiding over-fitting and improving the generalization performance, (3) decreasing required time for model training by reducing the number of features, (4) reducing the cost for collecting and storing data, and (5) gaining a deeper realization about the underlying processes that generated the data [8–10]. From the classification point of view, feature selection techniques can be divided into three groups of filter, wrapper and embedded methods [9]. Filter method, which is a learning-free technique, uses different statistical tests to determine the subset of features with the highest score obtained by an objective function [11,12]. Wrapper methods explore for the best feature subset in combination with a specific classifier model [13]. Wrapper methods typically give more accurate results, but they do more computations to search for the best features. Similar to wrapper methods, embedded methods [14] are assigned to a learning algorithm. They incorporate feature selection and the learning part of model in such a way that searching for an optimal subgroup of features is combined with the construction of classifier. The advantage of embedded methods is that they choose highly informative feature subsets for a specific model; moreover, they have less computational cost than wrapper methods [9,14].

Feature ranking is used in many feature selection methods as their principal or auxiliary selection mechanism. In this regard, features are ranked according to the relative score values computed by a criterion function, then a small number of top ranked features are selected as the most informative features. Feature ranking has several advantages, such as: its simplicity, scalability, and good empirical success for a variety of real-world applications [15,16].

Microarray data analysis and proteomic patterns exploration are important examples of feature ranking applications [17–19]. Feature ranking is conventionally performed by an individual criterion, which scores features according to their information content. Therefore, rank

[*] Corresponding author at: Department of Bioelectric and Biomedical Engineering, School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences (IUMS), Isfahan, Iran.
E-mail address: mr.sehhati@amt.mui.ac.ir (M. Sehhati).

https://doi.org/10.1016/j.jbi.2019.103213
Received 1 October 2018; Received in revised form 28 April 2019; Accepted 17 May 2019
Available online 23 May 2019
1532-0464/ © 2019 Elsevier Inc. All rights reserved.

of features depends on selected criterion and choosing an appropriate criterion will be crucial. Nguyen et al. proposed a novel method called modified analytic hierarchical process (AHP) to select prominent gene subsets for cancer classification using DNA microarray data [20–23]. Their method deals with five individual gene ranking methods: Entropy, receiver operating characteristic curve, signal to noise ratio, *t*-test, and Wilcoxon and combines them to get better performance in feature selection. Modified AHP yielded stable gene subsets that were informative features applied to different classification models. However, the integration procedure used by modified AHP is simplistic and do not consider statistical dependency among different ranking criteria. Here, we introduce a novel feature selection method that combines five feature ranking criteria, considering their relevance, by HMM. To the best of author's knowledge, this is the first presentation of HMM for feature selection in microarray gene expression profiles analysis.

The paper is organized in four sections. Section 2 describes our proposed method. Evaluation of the proposed method and the obtained results are expressed in Section 3. Finally, the paper is concluded in Section 4.

2. Methods

Many standard analytical techniques are inappropriate or computationally infeasible for analyzing high- throughput data. However, the dimension of data can be reduced by eliminating non-relevant and non-discriminative genes that do not participate in a specific biological phenomenon. The use of non-associated genes in data analysis enlarges the dimension of problem, increases the computational cost, and inserts misleading noise in the model. Thus, it is crucial to select a small set of associated genes, called informative genes, to have an acceptable performance in classification. The proposed method is based on fusion of different ranking criteria by HMM. This method assembles prominent discriminative features from different features ranking methods through the structure of HMM.

2.1. Feature ranking criteria

A feature-ranking criterion can be used to determine which available features are more appropriate for classification. Subsequently, features can be selected from the feature list ordered by the criterion function. In the following, some well-known feature ranking criteria used in our method are defined.

2.1.1. Two-sample t-test

The two-sample *t*-test is one of the prevalent filter methods for feature selection. The *t*-test method evaluates whether the means of two classes are statistically significant different from each other [24]. The *t*-test score is expressed by:

$$t = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} \tag{1}$$

where sample means of two classes are indicated by μ_1 and μ_2; σ_1 and σ_2 are the sample standard deviations; n_1 and n_2 are the sample sizes.

The *t-test* score is calculated on each feature by dividing the expression levels according to the class label and the feature with higher score value is more informative in contrast other features.

2.1.2. Entropy test

The relative entropy between two probability distributions on a random variable, is a measure of the distance between them [25]. The entropy score for each feature is calculated by the following relation:

$$e = \frac{1}{2}\left[\left(\frac{\sigma_1{}^2}{\sigma_2{}^2} + \frac{\sigma_2{}^2}{\sigma_1{}^2}\right) + \left(\frac{1}{\sigma_1{}^2} + \frac{1}{\sigma_2{}^2}\right)(\mu_1 - \mu_2)^2\right] \tag{2}$$

μ_1, σ_1, μ_2 and σ_2 are the mean and standard deviation of samples in class 1 and class 2, respectively. For each feature, the entropy score is calculated and features will be sorted according to their scores.

2.1.3. Receiver operating characteristic (ROC) curve

A receiver operating characteristic (ROC) curve is a graphical scheming of the true positive rate vs. the false positive rate for a binary classification model, when its decision threshold is varied. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two classes. The larger is the AUC, the less is the error of classification. In feature ranking application, features with the highest AUC are selected [26].

2.1.4. Wilcoxon method

The Wilcoxon signed rank sum test is used to test the null hypothesis that the median of a distribution is equal to a specific value {Nguyen, 2015 #164}[27]. Wilcoxon, which is a non-parametric test, can be used instead of the *t*-test to produce a null hypothesis in cases when the observation does not follow normal distribution.

The three-steps procedure of the Wilcoxon test is described below [28]:

1. Combine all observations from the two populations and sort them in the ascending order.

2. Calculate the Wilcoxon statistic by adding all the ranks related with the observations from the smaller population.

3. Finally, select the features whose p-values are smaller than the significance level threshold.

By employing the absolute values of the standardized Wilcoxon statistics as the feature scores, we can use the Wilcoxon test for feature selection.

2.1.5. Bhattacharyya distance

Bhattacharyya distance is generally used for measuring the similarity of two continuous or discrete probability distributions. The Bhattacharyya distance between two classes under the normal distribution can be calculated by the following [29]:

$$b = \frac{1}{4}ln\left(\frac{1}{4}\left(\frac{\sigma_1{}^2}{\sigma_2{}^2} + \frac{\sigma_2{}^2}{\sigma_1{}^2} + 2\right)\right) + \frac{1}{4}\left(\frac{(\mu_1 - \mu_2)^2}{\sigma_1{}^2 + \sigma_2{}^2}\right) \tag{3}$$

where sample means are indicated by μ_1 and μ_2 in class 1 and class 2, respectively and its corresponding, sample standard deviations are indicated by σ_1 and σ_2.

2.2. Markov chain rank aggregation

Andrey Markov introduced the Markov chain in 1906 and used the term “chain” for the first time when he was working on the theory of stochastic processes [30,31]. In mathematics generally, probability theory and statistics the term Markov property refers to random process characterized as memoryless: the next state depends only on the current state and do not depend on the sequence of events that happened before. Markov chain used in many applications and statistical modeling such as page rank, which was utilized for Google search engine. A Markov chain model is determined by a set of states; $S = \{1, 2, \dots, |S|\}$ and a non-negative stochastic (sum of each row is 1) matrix T of size $|S| \times |S|$ defines the probability of the systems' transitions from one state to another. To solve the steady-state transition probabilities, it is acceptable to either calculate the dominant eigenvector or use the power-iteration algorithm [32].

Markov chain rank aggregation is one of the most commonly pairwise comparison ranking methods [33]. The most popular example of Markov chain ranking method is ranking of web pages in Google search engine. In our case, the states of the chain correspond to the *N* features to be ranked and the transition probabilities are based on the positions of the features in each feature-ranking criterion. The transition matrix

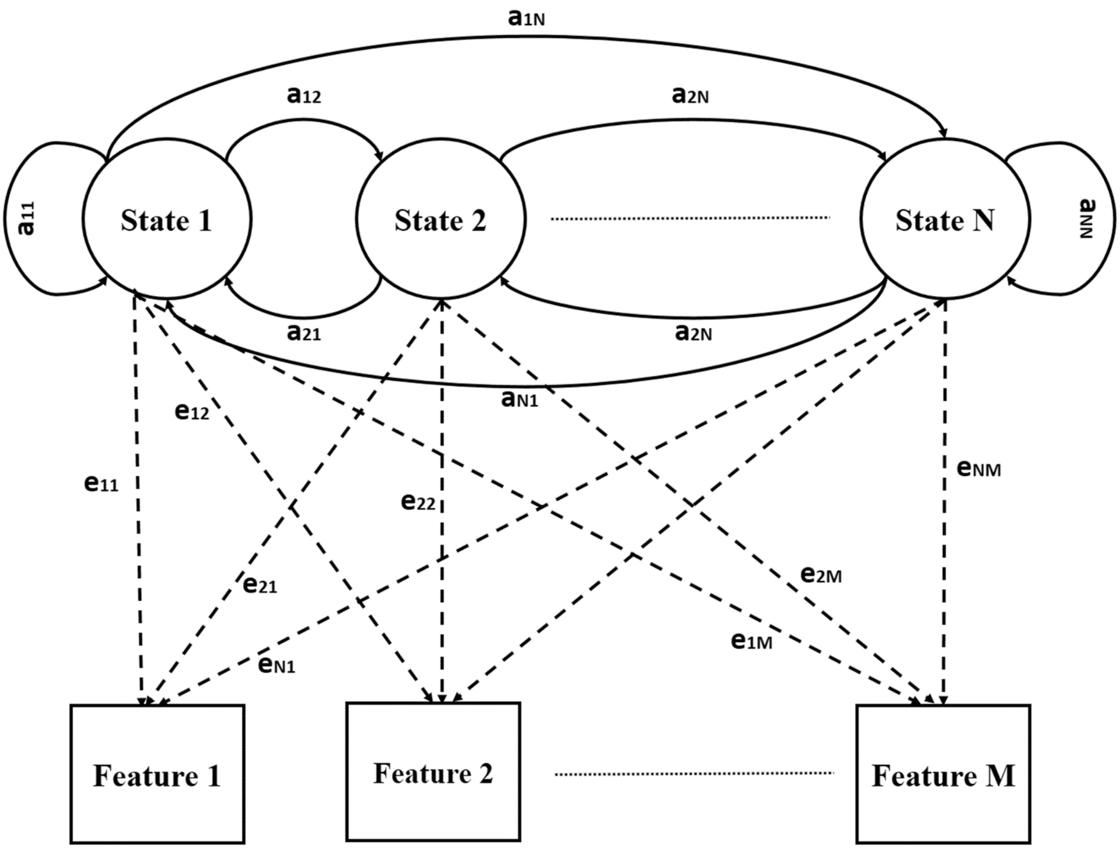

Fig. 1. HMM topology for proposed feature selection method.

of the Markov chain rank aggregation can be obtained by implementing the following instructions [32,34]:

1. Construct the set of states S corresponds to the list of all features to be ranked, i.e., the set of all features that ranks with each criterion $U = \{c_1, c_2, \dots, c_n\}$
2. For each pair of features *i* and *j* in *U*, let the component t^*_{ij} equal *1* if most criterion ranked *j* higher than *i*, and *0* otherwise. If features *i* and *j* are not directly compared, let $t^*_{ij} = t^*_{ji} = 0.5$
3. Describe the transition matrix $T = \{t_{ij}\}$ as fallow: for $i \neq j$ set t_{ij} to t^*_{ij} / $|U|$ and let $t_{ii} = 1\text{-}\Sigma_{j \neq i}\, t_{ij}.$
4. Multiplying each component by $(1 - \varepsilon)$ to make the transition matrix ergodic and then adding ε / $|U|$ to each component, where ε has a small, positive value.

After constructing the transition matrix, to find rating vector for feature ranking, either calculate the dominant eigenvector of transition matrix, or use the power method to obtain the steady-state probability vector.

2.3. Hidden Markov model

Baum et al. have introduced the HMM in the late 1960s [35–38]. HMM is a powerful statistical Markov model, which can be used, in a wide range of applications in which the system being modeled is supposed to be a Markov process with unobserved states. Each state of HMM has a probability distribution over the possible output symbols. Therefore, the sequence of generated symbols by an HMM gives some information about the hidden states.

Generally, HMM is determined by five elements of states, state probabilities, initial probabilities, transition probabilities and emission probabilities that are defined as following [39]:

1. The *N* states of the model are defined in S as:

$$S = \{S_1, \cdots, S_N\} \tag{4}$$

2. The *M* observation symbols per state are defined by:

$$V = \{\nu_1, \cdots, \nu_M\} \tag{5}$$

3. The state transition probability distribution, also called transition matrix $A = \{a_{ij}\}$, representing the probability of going from state S_i to state S_j:

$$a_{ij} = P\{q_{t+1} = S_j | q_t = S_i\}, \quad 1 \le i, j \le N \tag{6}$$

where q_t denotes the current state.

The transition probabilities should satisfy the following constraints:

$$a_{ij} \ge 0, 1 \le i, j \le N, \quad \sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N \tag{7}$$

4. The observation symbol probability distribution in each state, also called emission matrix $E = \{e_j\,(k)\}$ where $e_j\,(k)$ is the probability that symbol ν_k is emitted in state S_j .

$$e_j(k) = P\{o_t = \nu_k | q_t = j\}, \quad 1 \le j \le N, 1 \le k \le M \tag{8}$$

where ν_k represents the k^{th} observation symbol in the alphabet, and O_t the recent parameter vector.

The following constraints should be satisfied:

$$e_j(k) \ge 0, \quad 1 \le j \le N, \quad 1 \le k \le M \;\; and \;\; \sum_{j=1}^{N} e_j(k) = 1, \quad 1 \le j \le N \tag{9}$$

5. The initial state probability distribution $\pi = \{\pi_i\}$, representing

probabilities of states in t = 0.

$$\pi_i = p\{q_1 = i\}, \quad 1 \le i \le N \tag{10}$$

We indicate an HMM with parameter set λ = (A; E; π), which completely describes the model. These parameters are used to solve three well-known fundamental problems of HMMs [39]. First in the evaluation problem of HMM, forward or backward algorithm is used to calculate $P\{O|\lambda\}$, i.e. the probability that the given observation sequence $O=\{o_1, o_2, \ldots, o_T\}$ is generated by the model λ. Second in the decoding problem of HMM, Viterbi algorithm is applied to solve the second problem. Viterbi computes most likely state sequence associated with the given observation sequence and the model λ. Third in the learning problem of HMM, Baum Welch algorithm is used to train the HMM or to adjust the model parameters *(A; E; π)* in order to maximize $P\{O|\lambda\}$ given a sequence of observation and model λ.

In the proposed architecture, each hidden state of the HMM represents the probability that the best features are related to one of criterion and the feature rank (feature position) obtained by each criterion represents the corresponding sequence of observation. Fig. 1 shows the topology of the proposed structure, which illustrates states, observations and how they are connected to each other.

2.3.1. Transition matrix definition

According to the proposed architecture, shown in Fig. 1, the transition probabilities between states will be equal to the percentage of common high-ranked features observed between different criteria on the training data. Therefore, the transition matrix is constructed considering overlap among the first (high score) one percent of all features that ranked by 5 different ranking criteria in training data. As shown in Fig. 2, every component of transition matrix represents normalized number of common observations among different states of HMM, which are pairwise common features obtained by each criterion. F_i indicates one percent of high rank features selected by criterion *i* where *i* = *{t (t_test), e (entropy), r (ROC), w (Wilcoxon), b (Bhattacharyya)}*. Normalization was performed by dividing the number of common features between two different criterions by sum of all overlapping features between source criterion and other criteria. Thus, sum_t, sum_e, sum_r, sum_w and sum_b indicate sum of all overlapping feature for *t*-test, entropy, ROC, Wilcoxon and Bhattacharyya respectively as stated in the following:

$$sum_M = \sum\nolimits_{M \ne i} (F_M \cap F_i) \tag{11}$$

By this structure sum of every row of transition matrix will be equal to 1, so the probability of sample space or sum of probabilities is satisfied in the model. In this topology, the components on the main diagonal of transition matrix are equal to zero.

2.3.2. Emission matrix definition

Emission matrix represents probability distribution in each state (matrix row) and indicate the probability that feature *j* (matrix column) is emitted in state *i*, which is in concordance with each feature's rank obtained by each state. In this regard, feature with higher rank will be represented by a higher value in emission matrix. In order to satisfy probabilistic properties in HMM, we convert rank of all features for every criterion to a probabilistic score value by the following definition to construct the emission matrix (E);

$$T= \begin{array}{c|ccccc|} & \textbf{State 1} & \textbf{State 2} & \textbf{State 3} & \textbf{State 4} & \textbf{State 5} \\ \textbf{State 1} & 0 & \frac{(F_t \cap F_e)}{sum_t} & \frac{(F_t \cap F_r)}{sum_t} & \frac{(F_t \cap F_w)}{sum_t} & \frac{(F_t \cap F_b)}{sum_t} \\ \textbf{State 2} & \frac{(F_e \cap F_t)}{sum_e} & 0 & \frac{(F_e \cap F_r)}{sum_e} & \frac{(F_e \cap F_w)}{sum_e} & \frac{(F_e \cap F_b)}{sum_e} \\ \textbf{State 3} & \frac{(F_r \cap F_t)}{sum_r} & \frac{(F_r \cap F_e)}{sum_r} & 0 & \frac{(F_r \cap F_w)}{sum_r} & \frac{(F_r \cap F_b)}{sum_r} \\ \textbf{State 4} & \frac{(F_w \cap F_t)}{sum_w} & \frac{(F_w \cap F_e)}{sum_w} & \frac{(F_w \cap F_r)}{sum_w} & 0 & \frac{(F_w \cap F_b)}{sum_w} \\ \textbf{State 5} & \frac{(F_b \cap F_t)}{sum_b} & \frac{(F_b \cap F_e)}{sum_b} & \frac{(F_b \cap F_r)}{sum_b} & \frac{(F_b \cap F_w)}{sum_b} & 0 \end{array}$$

Fig. 2. Transition matrix structure for proposed HMM-based feature selection method.

$$E(i,j) = \frac{\alpha^{-R(i,j)}}{\sum_{k=1}^{M} \alpha^{-R(i,k)}}, \quad \alpha > 1 \tag{12}$$

where α is the base value of *E(i,j)*, *R(i,j)* indicates rank of j^{th} feature for i^{th} criterion, and *M* represents the number of features.

2.3.3. Learning

As mentioned in Section 2.2, Baum-Welch algorithm can be used to adjust the HMM parameters for the best representation of observations (feature rank obtained by each criterion in the training set) by the model. The Baum–Welch algorithm is a special case of Expectation-Maximization (EM) algorithm [40] and used to find the unknown parameters of an HMM. We describe Baum-Welch algorithm by defining several auxiliary variables [41]. The first one of these variables is:

$$\xi_t(i,j) = p\{q_t = i,\, q_{t+1}=j|O,\, \lambda\} \tag{13}$$

which can also be written as:

$$\xi_t(i,j) = \frac{p\{q_t = i,\, q_{t+1}=j,\, O|\lambda\}}{p\{O|\lambda\}} \tag{14}$$

The forward variable $\alpha_t(i)$ is defined as follows:

$$\alpha_t(i) = p\{O_1,\, O_2, \cdots, O_t,\, q_t=i|\lambda\} \tag{15}$$

where O_1, O_2 ,…,O_T are partial ranking sequences. The recursive relationships are as follow:

$$\alpha_{t+1}(j) = e_j(o_{t+1}) \sum_{i=1}^{N} \alpha_t(i) o_{ij},\ 1 \le j \le N,\ 1 \le t \le T-1 \tag{16}$$

$$\alpha_1(j) = \pi_j e_j(o_1), \quad 1 \le j \le N \tag{17}$$

The backward variable $\beta_t(i)$ can be defined similarly:

$$\beta_t(i) = p\{O_{t+1},\, O_{t+2}, \cdots, O_t|q_t = i,\, \lambda\} \tag{18}$$

If the current state is *i*, β_t *(i)* is the probability of the partial ranking sequence O_{T+1} ,O_{T+2} ,…,O_T. β_t *(i)* can also be computed by using the following recursive formula:

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} e_j(o_{t+1}), \quad 1 \le i \le N,\ 1 \le t \le T-1 \tag{19}$$

where

$$\beta_t(i) = 1, \quad 1 \le i \le N \tag{20}$$

We can now calculate the ξ_t *(i,j)* variable using forward and backward variables:

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} \beta_{t+1}(j) e_j(o_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} \beta_{t+1}(j) e_j(o_{t+1})} \tag{21}$$

Another variable is the *a posteriori* probability,

$$\gamma_t(i) = p\{q_t=i|O,\, \lambda\} \tag{22}$$

In forward and backward variables this can be indicated by (23).

$$\gamma_t(i) = \left[\frac{\alpha_t(i)\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)}\right] \tag{23}$$

The relationship between γ_t *(i)* and ξ_t *(i, j)* is given by (24).

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j), \quad 1 \le i \le N, 1 \le t \le M \tag{24}$$

Baum-Welch learning process is used to maximize $p\{O|\lambda\}$ and determine HMM parameters using iterative EM algorithm, by assuming a starting model $\lambda = (A,E,\pi)$ and calculating the forward and backward variables. Afterwards, the values of ξ and γ are calculated. The following equations are known as Baum-Welch re-estimation formulas, which are used to update the HMM parameters:

$$\bar{\pi}_i = \gamma_1(i), \quad 1 \le i \le N \tag{25}$$

$$\bar{a}_{ij} = \frac{Expectednumberoftransitionsfromstates_j tostates_i}{Expectednumberoftransitionsoutofstates_j} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)},$$
$$1 \le i \le N, \quad 1 \le j \le N \tag{26}$$

$$\bar{e}_j(k) = \frac{Expectednumberoftimesfeaturesoccursinstates_i}{Expectednumberoftransitionsoutofstates_j} = \frac{\sum_{\substack{t=1 \\ o_t=v_k}}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)},$$
$$1 \le j \le N, \quad 1 \le k \le M \tag{27}$$

We repeated the sequence of the above steps until 100 levels of convergence is achieved.

2.3.4. Multi-criteria feature ranking

After using Baum-Welch method and getting the HMM parameters, we consider the combination of values for the emission matrix elements as feature scores. In this regard, we added top score of all features (each column of emission matrix) for each state. It is expected that the larger the score, the better the feature will be. For every feature, we calculate the mean of the normalized values of emission matrix (column-wise) from all criteria and construct the scores as fallowing:

$$\text{Score}_j = \frac{1}{n}\sum_{i=1}^{n} e_{ij} \tag{28}$$

where e_{ij} denotes emission matrix component ($E = [e_{ij}]$) that estimated by Baum-Welch method, n represents the number of criteria and $Score_j$ indicates the score of individual feature j. After calculating the scores of all features, they are sorted in descending order and a new ranking for all features is obtained.

3. Experimental results

3.1. Evaluation metrics

Since the number of samples in our datasets is low, the leave one out cross validation (LOOCV) is used for evaluation. In each trial, LOOCV uses a single sample from the original data as the validation data, and the residual samples are used for training. This is repeated such that each sample in the data is used only once as the validation data. In other words, LOOCV is a special case of k-fold cross validation where k equals the number of samples in the data and each sample in the data is used exactly once for validation [42–44].

AUC (area under the receiver operating characteristic curve) is used to measure the effectiveness of classification. AUC is a widely used evaluation criterion to describe a diagnostic test especially for binary classification. To obtain reliable performance estimation and robust comparison among feature selection methods, large number of estimates are always preferred. Therefore, we increase the number of estimates by 20 times repeating of LOOCV cross-validation.

3.2. Experimental datasets

Three benchmark datasets used for experiments consist of diffuse large B-cell lymphomas (DLBCL) [45], leukemia cancer [46] and prostate [47]. In the first dataset, DLBCL and follicular lymphomas (FL) are two malignancies to be classified. The DLBCL dataset includes 7070 genes of 77 samples, where 58 samples are affected by DLBCL and the others have FL. The classification models are constructed using gene expression profiles to distinguish between these two lymphomas. The leukemia dataset comprises of acute lymphoblast leukemia (ALL) and acute myeloid leukemia (AML) samples from bone marrow and peripheral blood. This dataset contains 72 samples (47 ALL and 25 AML) where expressions are measured over 7,129 genes. In the third dataset, there exist 102 samples (50 normal and 52 prostate tissues samples) where each expression profile contains 12,533 genes. All the gene profile datasets are normalized by the quantile normalization technique [48].

3.3. Model parameters

3.3.1. Initialization and weighting

Identification of the best state for selecting the best features from a particular database is not straightforward and should take into account the nature of data. This task assigned to the proposed HMM model, however, initially the same level of importance was considered for each state. Hence, the value of 1/5 is assigned to each state for initial state probability distribution in HMM, because we used 5 criteria in this experiment.

3.3.2. α-parameter

α-parameter in (12) should be defined for construction of emission matrix. If we select α value close to 1, the score values of top rank features would be low and intervals between score values are shrunk. For greater values (e.g. > 4), the scores values for the high rank features would be greater and intervals between score values is increased, which may lead to missing middle and low rank features due to very low score values (approximately zero). The optimum range of values for α-parameter is between 1 and 4 which can be tuned for different datasets.

3.3.3. Number of features

Feature selection chooses a less number of features from the initial feature set without changing the original properties of features [9]. To make an appropriate comparison between the proposed HMM feature selection model and other feature selection methods (*t*-test, entropy, ROC, Wilcoxon, Bhattacharyya distance and Markov chain), we applied different number of features (5, 10, 15 and 20) to classifier. For evaluating the effect of number of features on the classification performance, we used k-nearest neighbors (kNN) and support vector machine (SVM) as classifiers that are widely used in gene expression classification [11,49]. The number of nearest neighbors in the kNN is chosen to be k = 5, and the linear kernel function is used for the SVM in all experiments.

3.4. Results and discussion

Classification results on three datasets (i.e., DLBCL, leukemia, and prostate) are demonstrated and discussed in this section. Fig. 3 shows the results on DLBCL dataset that exhibits a higher performance of HMM compared to other feature selection approaches using different number of selected features. The mean and standard deviation across replicates of AUC for different number of features are 94.50 ± 0.29, 91.12 ± 1.59 and 89.09 ± 1.52 for the proposed method, Wilcoxon and Markov chain respectively. According to Fig. 3a and Fig. 3b, the Wilcoxon method is in second place among all evaluated methods. Even though the Markov chain is a common rank aggregation method, yielded poor performance.

Fig. 4 illustrates the results obtained on leukemia dataset in which the proposed method reached the top AUC among all methods. Fig. 4a demonstrates that feeding extracted features by HMM to SVM results in AUC of 100% over 5, 10 and 15 selected features in leukemia dataset.

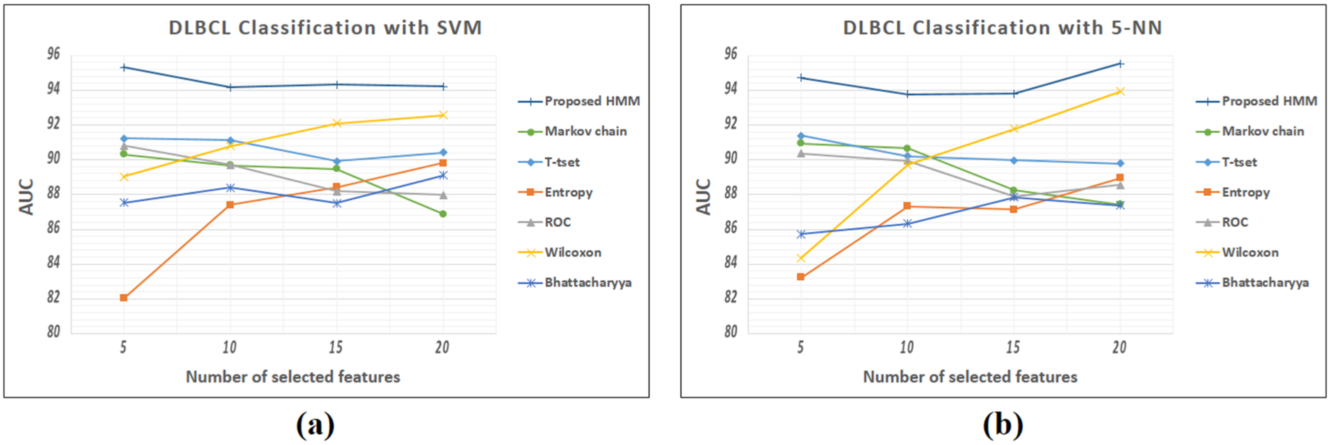

Fig. 3. Classification results for DLBCL datasets over different number of selected features: (a) DLBCL classification with SVM, (b) DLBCL classification with 5-NN.

Fig. 4b represents that ROC obtained the best performance over 5 selected features using 5-NN classifier. Wilcoxon method achieved the worst performance in leukemia dataset. Also our HMM based method has an AUC of 99.86 ± 0.27 across different selected features in leukemia dataset using SVM as classifier. The corresponding values for the ROC and Wilcoxon are 99.29 ± 0.54 and 87.80 ± 1.51 respectively.

Classification results of prostate dataset are illustrated in Fig. 5 and HMM exhibit best performance for various numbers of selected features. Fig. 5a demonstrates that *t*-test, Bhattacharyya distance and Markov chain have same performance over different numbers of features. Moreover according to Fig. 5a the best performance of proposed HMM method is obtained by selecting 10 features in SVM classifier. Fig. 5b illustrate that entropy reached the worst performance in prostate dataset by 5-NN classifier. Also according to Fig. 5b, HMM has the best performance by 15 features. Moreover, it can be observed that in the selected range of feature numbers, the variation of AUC is not noticeable in our model. The AUC across different selected features is about 92.10 ± 1.32 for the proposed method in the prostate dataset using the SVM. These values for the Wilcoxon and ROC, which have second and third place, are 90.34 ± 1.54 and 89.18 ± 2.078, respectively.

In addition, we conclude from Fig. 3, Fig. 4 and Fig. 5 that proposed HMM yields best results over all number of features and slope of changes in AUC results is the lowest for HMM, which proves the stability and less variation of our model across different selected features. Also using lower number of features, HMM obtained better performance in contrast to other methods.

Every feature selection method may have its own best performance by a different number of features in each dataset. However, our HMM-based method showed less variation in different datasets using different number of features. The proposed multi-criteria decision making method incorporates patterns and advantages of every constructive method. In this regard, transition matrix constructed by considering overlap between different method's outcomes, represents the association among individual states. This procedure is more realistic than the technique presented by Nguyen [20–23], modified AHP, which considers equal level of importance for each criterion. Furthermore, the modified AHP reached the maximum AUC of 94.77 ± 3.98, 88.62 ± 3.13 and 88.75 ± 2.99 in the DLBCL, leukemia, and prostate datasets respectively by the SVM classifier using five selected genes. The corresponding values for the proposed HMM-based method are 95.22 ± 1.56, 100.0 ± 0.00 and 91.73 ± 0.48. Therefore, the proposed method showed better performance and robustness compared to the modified AHP. This is a fair comparison because the utilized datasets are the same in these researches. Moreover, the proposed HMM-based method compared with Markov chain rank aggregation in a way that both approach was used a combination of similar methods and use Markov property in their structures. Markov chain rank aggregation uses pairwise comparisons between the features to determine rankings and operates based on voting between ranking methods. While our HMM-based approach uses overlapping among the first one percent of all features that ranked by ranking methods in the form of transition probability between hidden states. However, the results obtained from Fig. 3, Fig. 4 and Fig. 5 show that the HMM-based method has a much better performance than the Markov chain rank aggregation across different number of selected features. Also, AUC results of the Markov

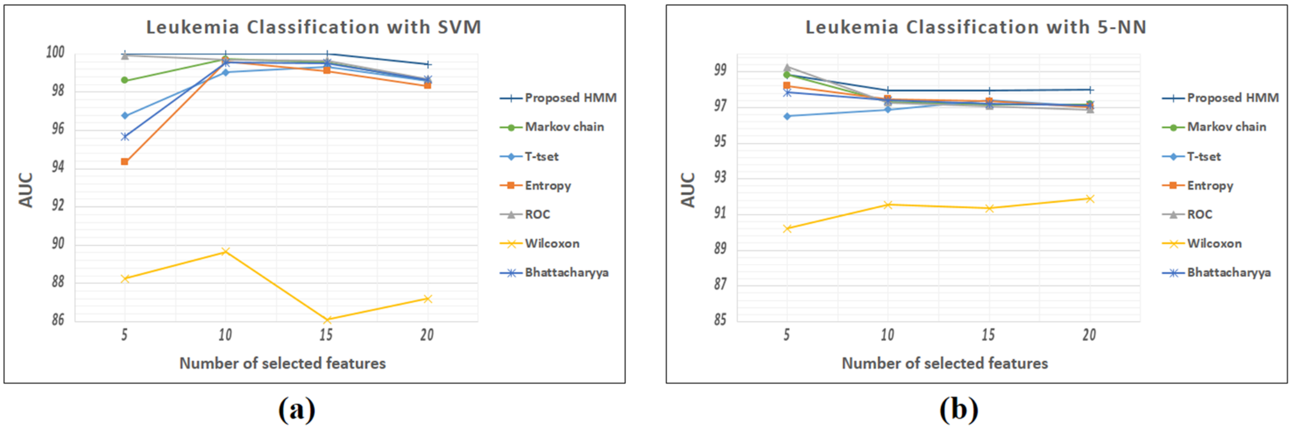

Fig. 4. Classification results for leukemia datasets over different number of selected features: (a) Leukemia classification with SVM, (b) Leukemia classification with 5-NN.

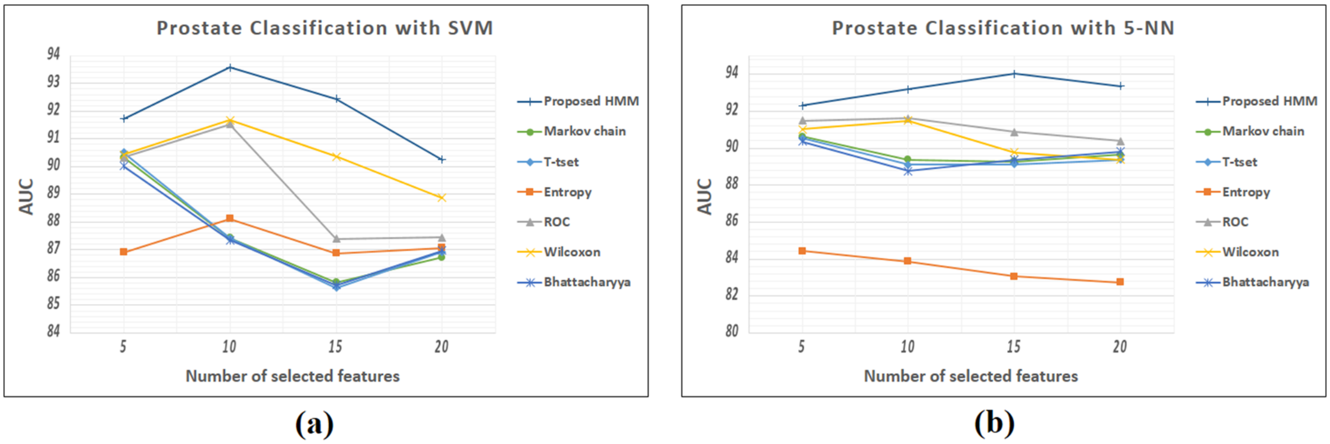

Fig. 5. Classification results for prostate datasets over different number of selected features: (a) Prostate classification with SVM, (b) Prostate classification with 5-NN.

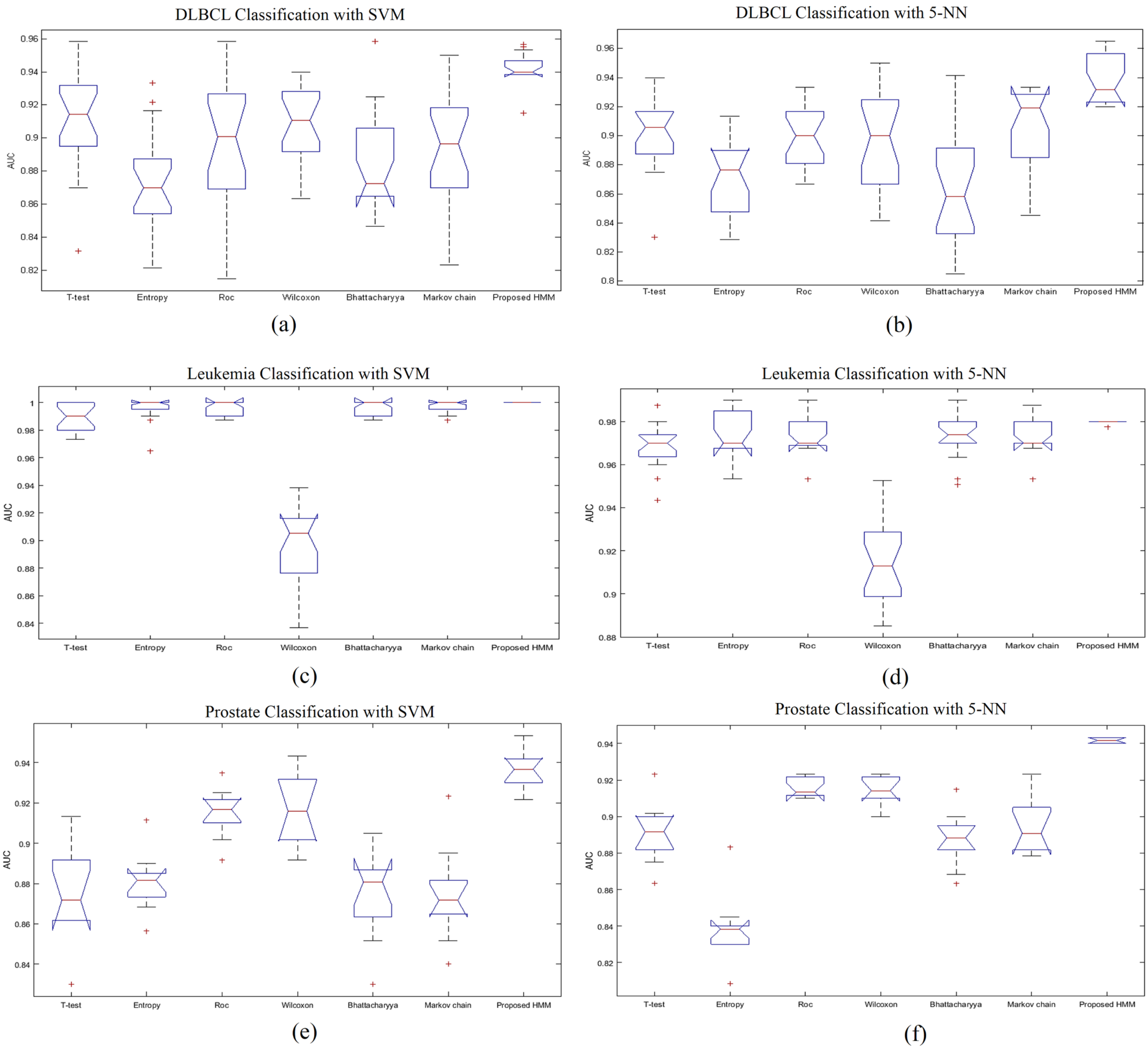

Fig. 6. Boxplot comparisons among feature selection methods over 10 selected features: (a) DLBCL classification with SVM, (b) DLBCL classification with 5-NN, (c) Leukemia classification with SVM, (d) Leukemia classification with 5-NN, (e) Prostate classification with SVM, (f) Prostate classification with 5-NN.

chain showed much more variation than our HMM-based method across different selected features, which reflect the better stability of the proposed method.

Fig. 6 shows boxplots illustrating the performance comparisons in the form of AUC among HMM and other feature selectors over 10 features. We observed a noticeable dominance of the HMM against other feature selection methods in all three datasets. Furthermore, small interquartile varieties of the HMM boxes compared to other methods display smaller standard deviations and demonstrates the robustness and greater stability of HMM compared to other feature selection methods.

4. Conclusions

In this paper, we presented a novel feature selection method based on HMM for cancer classification using gene expression data. This method was developed by combining five different feature ranking methods including: *t*-test, entropy, ROC, Wilcoxon test and Bhattacharyya distance in the topology of HMM. Our experiments implemented on three benchmark datasets and for better estimation LOOCV cross-validation was repeated 20 times. Classification performance was evaluated by AUC evaluation metric. Classification results demonstrated that HMM yields the best performance compared to five mentioned feature ranking methods and Markov chain rank aggregation method. Moreover, smaller standard deviation results proved that HMM was the most robust method compared to other feature selection methods. It also provided stable results over different number of selected features. In the future research, we would like to extend our HMM-based method using more feature selection methods to get better results.

Declaration of Competing Interest

None.

Acknowledgement

This research was financially supported by the Isfahan University of Medical Sciences, Isfahan, Iran (Grant no. 397018).

References

[1] A.R. Webb, Statistical Pattern Recognition, John Wiley & Sons, 2003.
[2] I. Jolliffe, Principal Component Analysis, Wiley Online Library, 2005.
[3] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[4] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2005) 1226–1238.
[5] H.-L. Wei, S.A. Billings, Feature subset selection and ranking for data dimensionality reduction, IEEE Trans. Pattern Anal. Mach. Intell. 29 (2007).
[6] H. Zeng, Y-m. Cheung, Feature selection and kernel learning for local learning-based clustering, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2011) 1532–1547.
[7] N. Kwak, C.-H. Choi, Input feature selection by mutual information based on Parzen window, IEEE Trans. Pattern Anal. Mach. Intell. 24 (2002) 1667–1671.
[8] T. Jirapech-Umpai, S. Aitken, Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes, BMC Bioinf. 6 (2005) 148.
[9] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (2007) 2507–2517.
[10] A. Jović, K. Brkić, N. Bogunović, A review of feature selection methods with applications, Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015 38th International Convention on: IEEE, 2015, pp. 1200–1205.
[11] M.R. Sehhati, A.M. Dehnavi, H. Rabbani, S.H. Javanmard, Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence, J. Med. Signals Sensors 3 (2013) 87.
[12] N. Sánchez-Maroño, A. Alonso-Betanzos, M. Tombilla-Sanromán, Filter methods for feature selection–a comparative study, Intell. Data Eng. Automated Learn.-IDEAL 2007 (2007) 178–187.
[13] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.
[14] T. Lal, O. Chapelle, J. Weston, A. Elisseeff, Embedded methods, Feature Extract. (2006) 137–165.
[15] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.
[16] W. Yan, Fusion in multi-criterion feature ranking, Information Fusion, 2007 10th International Conference on: IEEE, 2007, pp. 1–6.
[17] E.P. Xing, M.I. Jordan, R.M. Karp, Feature selection for high-dimensional genomic microarray data, ICML2001. pp. 601–608.
[18] Y. Wang, I.V. Tetko, M.A. Hall, E. Frank, A. Facius, K.F. Mayer, et al., Gene selection from microarray data for cancer classification—a machine learning approach, Comput. Biol. Chem. 29 (2005) 37–46.
[19] H. Liu, J. Li, L. Wong, A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, Genome Informat. 13 (2002) 51–60.
[20] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, A novel aggregate gene selection method for microarray data classification, Pattern Recogn. Lett. 60 (2015) 16–23.
[21] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, Hidden Markov models for cancer classification using gene expression profiles, Inf. Sci. 316 (2015) 293–307.
[22] T. Nguyen, S. Nahavandi, Modified AHP for gene selection and cancer classification using type-2 fuzzy logic, IEEE Trans. Fuzzy Syst. 24 (2016) 273–287.
[23] T. Nguyen, A. Khosravi, D. Creighton, S. Nahavandi, Hierarchical gene selection and genetic fuzzy system for cancer microarray data classification, PLoS One 10 (2015) e0120364.
[24] M. Alshalalfah, R. Alhajj, Cancer class prediction: two stage clustering approach to identify informative genes, Intell. Data Anal. 13 (2009) 671–686.
[25] S. Cateni, V. Colla, M. Vannucci, A hybrid feature selection method for classification purposes, Modelling Symposium (EMS), 2014 European: IEEE, 2014, pp. 39–44.
[26] X-w. Chen, M. Wasikowski, Fast: a roc-based feature selection metric for small samples and imbalanced data classification problems, Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 124–132.
[27] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bull. 1 (1945) 80–83.
[28] L. Deng, J. Pei, J. Ma, D.L. Lee, A rank sum test method for informative gene discovery, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2004, pp. 410–419.
[29] E. Choi, C. Lee, Feature extraction based on the Bhattacharyya distance, Pattern Recogn. 36 (2003) 1703–1709.
[30] N.J. Nilsson, Principles of artificial intelligence, Morgan Kaufmann (2014).
[31] M.H. Davis, Markov Models & Optimization, Routledge, 2018.
[32] M.E. Renda, U. Straccia, Web metasearch: rank vs. score based rank aggregation methods, Proceedings of the 2003 ACM Symposium on Applied Computing, ACM, 2003, pp. 841–846.
[33] C. Dwork, R. Kumar, M. Naor, D. Sivakumar, Rank aggregation methods for the web, Proceedings of the 10th International Conference on World Wide Web, ACM, 2001, pp. 613–622.
[34] R.P. DeConde, S. Hawley, S. Falcon, N. Clegg, B. Knudsen, R. Etzioni, Combining results of microarray experiments: a rank aggregation approach, Statist. Appl. Genetics Mol. Biol. 5 (2006).
[35] L.E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat. 37 (1966) 1554–1563.
[36] L.E. Baum, J.A. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, Bull. Am. Math. Soc. 73 (1967) 360–363.
[37] L.E. Baum, T. Petrie, G. Soules, N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, Ann. Math. Stat. 41 (1970) 164–171.
[38] L.E. Baum, An inequality and associated maximization thechnique in statistical estimation for probabilistic functions of Markov process, Inequalities 3 (1972) 1–8.
[39] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (1989) 257–286.
[40] S.R. Eddy, Multiple alignment using hidden Markov models, ISMB (1995) 114–120.
[41] J.A. Bilmes, A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, Int. Comput. Sci. Instit. 4 (1998) 126.
[42] T.-T. Wong, Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation, Pattern Recogn. 48 (2015) 2839–2846.
[43] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinformat. Comput. Biol. 3 (2005) 185–205.
[44] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (2001) 673.
[45] S. Monti, K.J. Savage, J.L. Kutok, F. Feuerhake, P. Kurtin, M. Mihm, et al., Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response, Blood 105 (2005) 1851–1861.
[46] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[47] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, et al., Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2002) 203–209.
[48] B.M. Bolstad, R.A. Irizarry, M. Åstrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2003) 185–193.
[49] M. Sehhati, A. Mehridehnavi, H. Rabbani, M. Pourhossein, Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information, IEEE/ACM Trans. Comput. Biol. Bioinf. 12 (2015) 1440–1448.