Information Sciences 316 (2015) 293–307

Contents lists available at ScienceDirect

Information Sciences

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

Hidden Markov models for cancer classification using gene expression profiles

Thanh Nguyen *, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi

Centre for Intelligent Systems Research (CISR), Deakin University, Waurn Ponds Campus, Victoria 3216, Australia

ARTICLE INFO

Article history:
Received 8 October 2014
Received in revised form 24 December 2014
Accepted 9 April 2015
Available online 16 April 2015

Keywords:
Analytic hierarchy process
Hidden Markov model
Cancer classification
DNA microarray
Gene expression profile
Gene selection

ABSTRACT

This paper introduces an approach to cancer classification through gene expression profiles by designing supervised learning hidden Markov models (HMMs). Gene expression of each tumor type is modelled by an HMM, which maximizes the likelihood of the data. Prominent discriminant genes are selected by a novel method based on a modification of the analytic hierarchy process (AHP). Unlike conventional AHP, the modified AHP allows to process quantitative factors that are ranking outcomes of individual gene selection methods including *t*-test, entropy, receiver operating characteristic curve, Wilcoxon test and signal to noise ratio. The modified AHP aggregates ranking results of individual gene selection methods to form stable and robust gene subsets. Experimental results demonstrate the performance dominance of the HMM approach against six comparable classifiers. Results also show that gene subsets generated by modified AHP lead to greater accuracy and stability compared to competing gene selection methods, i.e. information gain, symmetrical uncertainty, Bhattacharyya distance, and ReliefF. The modified AHP improves the classification performance not only of the HMM but also of all other classifiers. Accordingly, the proposed combination between the modified AHP and HMM is a powerful tool for cancer classification and useful as a real clinical decision support system for medical practitioners.

1. Introduction

DNA microarray is a collection of microscopic spots attached to a solid surface to measure the expression levels of genes. This technology enables researchers to study simultaneously a large number of genes (approximately 21,000 genes in the human genome). The cancer diagnosis using gene expression profiles therefore has been tremendously advanced. Cancer is basically a group of diseases when relevant genes stop functioning properly. In order to better diagnose, understand, and treat cancer, it is important to investigate which of the genes in cancer cells are working abnormally. Once subsets of differentially expressed genes are identified, classification techniques may be employed to distinguish cancer cells and normal cells.

Khan et al. [16] introduced a method for classifying cancers to specific diagnostic categories based on their gene expression signatures using artificial neural networks. The authors developed a stringent quality filter to include only the genes for which there were good measurements for all samples. Likewise, a procedure for multiclass cancer classification using

* Corresponding author. Tel.: +61 3 52278281; fax: +61 3 52271046.
E-mail address: thanh.nguyen@deakin.edu.au (T. Nguyen).

http://dx.doi.org/10.1016/j.ins.2015.04.012
0020-0255/

multivariate partial least squares (PLS) dimension reduction combined with logistic discrimination or quadratic discriminant analysis classifiers was suggested in Nguyen and Rocke [19].

In another approach, a process that decomposes multiclass ranking statistics into class-specific statistics and uses Pareto-front analysis for gene selection was recommended in Rajapakse and Mundra [23]. You et al. [32] implemented a local dimension reduction algorithm TotalPLS based on PLS to select prominent genes for classification. Alternatively, a hybrid approach that embeds the Markov blanket with the harmony search algorithm for gene selection was suggested by Shreem et al. [25]. The procedure works well on selected genes with higher correlation coefficients based on symmetrical uncertainty.

More recently, Sun et al. [28] used kernel method to discover inherent nonlinear correlations among genes as well as between gene and target class. An iterative PLS algorithm based on backward variable elimination through the "variable influence on projection" statistic for gene selection and classification was initiated in Burguillo et al. [7]. Chen et al. [9] on the other hand utilized particle swarm optimization integrated with a decision tree to analyse gene expression data.

For evaluating a cancer classification approach, in addition to the predictive ability of gene subsets and classifiers, two other important aspects that need to be considered are the stability and computational costs. This paper introduces a hybrid method that combines a gene selector by modified analytic hierarchy process (AHP) and a classifier by hidden Markov models (HMMs). Accordingly, the contribution of this paper is twofold. First, it proposes a substantial modification to the conventional AHP to account for quantitative criteria that are statistical ranking results of five individual filter methods. Second, a supervised classifier is designed exploiting an underlying HMM that takes genes selected by AHP as inputs.

Traditional AHP often deals with qualitative factors that are derived from experts. Given that the number of genes in microarray data are at around tens of thousands and the gene knowledge available to experts is always limited, completion of assessments of various genes with respect to various criteria is not always a practical proposition. We therefore propose a substantial modification to AHP for gene selection. The modified AHP is able to quantitatively integrate statistical outcomes of individual gene ranking methods via an objective ranking procedure without consulting to possibly biased and inadequate expert knowledge. Through rigorous experiments, we show that the modified AHP yields gene subsets that lead to a classification stability at low computational cost without sacrificing the accuracy.

On the other hand, HMMs are designed following a supervised learning approach so that they are capable of realizing knowledge available from cancer training data. Cancer often develops through different stages. These stages resemble the state transition of HMMs. In addition, the modularity characteristic of HMMs allows them to be combined into larger ones where each HMM is individually trained for each cancer data class. Given a new sample, trained HMMs can predict whether it is from a cancer or normal cell. To our best knowledge, this is the first application of HMMs as a classifier for cancer classification using gene expression profiles. Through this study, we examine and compare performance of HMMs with classification methods frequently applied in literature. Experiments are conducted using four microarray datasets to make sure conclusions driven out of this study are valid and general.

Details of the HMM approach implemented as a classifier are described in Section 3. Before that, Section 2 presents a background of gene selection and the modified AHP method. Experimental results are presented and discussed in Section 4, followed by conclusions in Section 5.

2. Gene selection methods

Microarray data are commonly assembled with the number of genes much larger than the number of samples [5]. Standard techniques therefore find inappropriate or computationally infeasible in analysing such data. Not all of the thousands of genes are discriminative and needed for classification. Most genes are not relevant to the cancer development and do not affect the classification performance. Taking such genes into account enlarges the dimension of the problem, leads to computational burden, and presents unnecessary noise in the classification process. Therefore it is essential to select a small number of genes, called informative genes, which can suffice for good classification. However, the best subset of genes is usually unknown [31].

Common gene selection approaches are filter and wrapper methods. Filter methods rank all features in terms of their goodness using the relation of each single gene with the class label based on a univariate scoring metric. The top ranked genes are chosen before classification techniques are executed. In contrast, wrapper methods require the gene selection technique to combine with a classifier to evaluate classification performance of each gene subset. The optimal subset of genes is identified based on the ranking of performance derived from implementing the classifier on all found subsets. The filter procedure is unable to measure the relationship among genes whilst the wrapper approach requires a great computational expense.

In this paper, to enhance the robustness and stability of microarray data classification, we introduce a novel gene selection method based on a modification of the AHP. The idea behind this approach is to incorporate prominent discriminant genes from different gene selection ranking methods through a systematic hierarchy.

The next subsections scrutinize background of common gene selection filter methods, which are followed by our proposal. The following gene selection methods rank genes via scoring metrics, which are statistic tests based on two data samples in the binary classification problem. The sample means are denoted as μ_1 and μ_2, whereas σ_1 and σ_2 are the sample standard deviations, and n_1 and n_2 are the sample sizes.

2.1. Two-sample t-test

The two-sample t-test is a parametric hypothesis test that is applied to compare whether the average difference between two independent data samples is really significant. The test statistic is expressed by:

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

In the application of t-test for gene selection, the test is performed on each gene by separating the expression levels based on the class variable. The absolute value of t is used to evaluate the significance among genes. The higher the absolute value, the more important is the gene.

2.2. Entropy test

Relative entropy, also known as divergence, is a test assuming classes are normally distributed. The entropy score for each gene is computed using the following expression [29]:

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

After the computation is accomplished for every gene, genes with the greatest entropy scores are selected to serve as inputs to the classification techniques.

2.3. Receiver operating characteristic (ROC) curve

Denote the distribution functions of X in the two populations as $F_1(x)$ and $F_2(x)$. The tail functions are specified respectively $T_i(x) = 1 - F_i(x),\ i = 1, 2$. The ROC is given as follows:

$$\mathrm{ROC}(t) = T_1\left(T_2^{-1}(t)\right), \quad t \in (0, 1) \tag{3}$$

and the area under the curve (AUC) is computed by:

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)dt \tag{4}$$

The larger the AUC, the less is the overlap of the classes. Genes with the greatest AUC therefore are chosen to form a gene set.

2.4. Wilcoxon method

The Wilcoxon rank sum test is equivalent to the Mann–Whitney U-test, which is a test for equality of population locations (medians). The null hypothesis is that two populations enclose identical distribution functions whereas the alternative hypothesis refers to the case two distributions differ regarding the medians. The normality assumption regarding the differences between the two samples is not required. That is why this test is used instead of the two-sample t-test in many applications when the normality assumption is concerned.

The main steps of the Wilcoxon test [12] are summarized below:

(1) Assemble all observations of the two populations and rank them in the ascending order.
(2) The Wilcoxon statistic is the sum of all of the ranks associated with the observations from the smaller group.
(3) The hypothesis decision is made based on the p-value, which is found from the Wilcoxon rank sum distribution table.

In the applications of the Wilcoxon test for gene selection, the absolute values of the standardized Wilcoxon statistics are utilized to rank genes.

2.5. Signal to noise ratio (SNR)

SNR defines the relative class separation metric by:

$$\mathrm{SNR}(f_i, c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2} \tag{5}$$

where c is the class vector, f_i is the ith feature vector. By treating each gene as a feature, we transform the SNR for feature selection to gene selection problem for microarray data classification.

SNR implies that the distance between the means of two classes is a measure for separation. Furthermore, the small standard deviation favors the separation between classes. The distance between mean values is thus normalized by the standard deviation of the classes [15].

2.6. A novel gene selection by modified AHP

Each of the above criteria can be employed to derive the ranking of genes and then to select greatest ranking genes for classification methods. The confidence of using a single criterion for selecting genes is not always achieved. Considering which criterion should be used is uncertain. This question inspires an idea of taking into account the ranking results of all criteria in evaluating genes. Through this way, salient genes of each criterion would be systematically assembled to form the most informative and stable gene subsets for classification. Furthermore, it is a difficult practice to combine ranking results of all criteria because the statistical ranges of criteria are different. The criterion generating a greater range of values would dominate those with a lower range. In order to avoid this problem, we utilize AHP in evaluating genes. The AHP deployment however commonly deals with qualitative criteria where their evaluations are derived from experts. Nevertheless, experts' knowledge is often limited, particularly when the problem being solved is conducted on a wide number of criteria referring to various knowledge areas. This advocates the use of quantitative criteria in the AHP. The following presents a novel proposal related to a vis-à-vis a ranking procedure utilizing quantitative criteria in AHP for the gene selection problem. The criteria used herein are the five test statistics, i.e. *t*-test, entropy, ROC, Wilcoxon, and SNR.

The AHP method as broadly applied in complex multi-criteria decision making is often performed with a tree structure of criteria and sub-criteria [20]. Due to the nature of the criteria selected here, the tree structure has three levels of hierarchies as illustrated in Fig. 1.

Five criteria are considered simultaneously during the AHP implementation. The five criteria are all quantitative so that we can intuitively put actual figures of these criteria into elements of the pairwise ranking matrix. This however would distort the matrix relative to other matrices describing assessments and judgements with respect to other criteria. Conventional applications of hierarchical analysis usually draw on the Saaty rating scale [1,9] and rough ratios, e.g. 1, 3, 5, 7, 9 to build pairwise comparison matrices [24]. In this research, we propose the scale [1,10] for ranking importance or significance of a gene compared with other genes. This scale will be applied to all criteria in the AHP application.

Suppose $X = (x_{ij})$ is the $n \times n$-dimension pairwise judgement matrix in which each element x_{ij} represents the relative importance of gene i over gene j with respect to a determined criterion, n is the number of genes. The reciprocal characteristic induces the following constraints

$$x_{ij} = 1/x_{ji}, \quad \forall i \neq j, i, j \in [1, n] \tag{6}$$

$$x_{ii} = 1, \quad \forall i \in [1, n] \tag{7}$$

If gene i is absolutely more informative than gene j, then we have $x_{ij} = 10$. Accordingly gene j must be absolutely less important than gene i and $x_{ji} = 1/10$. Where $x_{ij} = 1$, this indicates that two genes are equally informative. The higher the value of $x_{ij} \in [1, 10]$, the more important the gene i is in comparing with gene j. Element x_{ij} that is greater than 1 is called a superior element. Otherwise x_{ji} is called an inferior element as it is smaller than 1.

Let us define distance d_{ij} between two genes i and j with respect to a given criterion (i.e. *t*-test, entropy, ROC, Wilcoxon or SNR) by the absolute value of the subtraction between two statistics c_i and c_j of two genes.

$$d_{ij} = abs(c_i - c_j) \rightarrow d_{ij} = d_{ji} \tag{8}$$

For all criteria, the higher the statistic, the more important the gene is. The procedure to acquire elements of comparison reciprocal matrices is described below where $c_{\max}$ is the maximum distance of genes regarding the given criterion, $c_{\max} = \max(d_{ij}),\ \forall i, j \in [1, n]$, and c is a temporary variable.

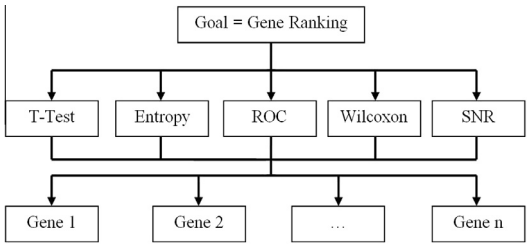

Fig. 1. The hierarchy of factors for gene selection by AHP.

Ranking Procedure:

FOR a pair of two genes i and j

$$c = (d_{ij} * 9)/(c_{\max}) + 1 = (abs(c_i - c_j) * 9)/(c_{\max}) + 1 \tag{9}$$

IF $(c_i \geqslant c_j)$ THEN $x_{ij} = c$ ELSE $x_{ij} = 1/c$ END IF
END FOR

The expressions of x_{ij} ensure that superior elements of the judgment matrices will be distributed in the interval [1,10]. Through calculations of the quantitative ranking method, the superior ratios are allowed to be real numbers within [1,10] so that they can reflect more rigorously the significance in judgements compared with the original Saaty rating scale. For example, consider four quantitative criteria A, B, C, and D with respective values 0.9, 1.3, 8.7, and 9.2. According to the Saaty rating scale, criteria B and A (D and C) are considered "equally important" and the ratios x_{BA} and x_{DC} will be equally assigned to 1: $x_{BA} = x_{DC} = 1$. Obviously, the difference between B and A (or D and C), though small, is neglected. However, with our ranking method, the ratios x_{BA} and x_{DC} are assigned more precisely and differently $1.43 = x_{BA} \neq x_{DC} = 1.54$. Likewise, in the Saaty rating scale, criterion C is considered absolutely more important than criterion A and B, and the ratio x_{CA} and x_{CB} are both assigned 9. In our scale, the ratio x_{CA} and x_{CB} will be assigned differently 9.46 and 9.02 respectively. Hence the "absolute importance" judgement is relaxed and replaced by more rigorous judgements with different real numbers 9.46 and 9.02 rather than the same rough number 9 for both x_{CA} and x_{CB}.

After comparison matrices are constructed, hierarchical analysis calculates eigenvectors that represent ranking scores of genes. Denote the comparison matrix as $X = \{x_{ij}\}$ where x_{ij} is the comparison between gene i and gene j, sum of values of elements are calculated as

$$S_j = \sum_{i=1}^{n} x_{ij} \tag{10}$$

then the elements of the eigenvector $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^T$ of X are estimated by:

$$\epsilon_i = \frac{1}{n} \sum_{j=1}^{n} \frac{x_i j}{S_j} \tag{11}$$

It is clear that $\sum_{i=1}^{n} \epsilon_i = 1$. While applying the AHP, the matrix is required to be consistent and hence its elements must be transitive, that is $x_{ik} = x_{ij} x_{jk}$. To verify the consistency of the comparison matrix X, the Consistency Index (CI) and Consistency Ratio (CR) based on large samples of matrices of purely random judgements are calculated. Let λ be an eigenvalue of the square matrix X, we have that $X\epsilon = \lambda\epsilon$. Then CI and CR are computed as follows:

$$\mathrm{CI} = \frac{\lambda_{\max} - n}{n - 1} \tag{12}$$

$$\mathrm{CR} = \frac{\mathrm{CI}}{\mathrm{RI}} \tag{13}$$

where $\lambda_{\max} = \max(\lambda_i),\ \forall i \in [1, n];\ \lambda_i = \frac{1}{\epsilon_i}[x_{i1}, \ldots, x_{in}][\epsilon_1, \ldots, \epsilon_n]^T$ and RI is the random consistency index that can be calculated using randomly generated reciprocal matrices.

CR should not exceed 0.1 if the set of judgements is consistent although CRs of more than 0.1 (but not too much more) sometimes have to be accepted in practice. CR equal to 0 implies the judgements are perfectly consistent.

Once calculations for five criteria are completed, we obtain the so-called option performance matrix consisting of five eigenvectors that has the form shown in Table 1.

Finally, the ranking of genes is the multiplication of the performance matrix and the vector representing the important weight of every criterion. The weight vector can be obtained by evaluating the significance of each criterion regarding the goal using the same procedure as described above. However, to avoid a bias judgement, we consider five criteria having an equal significance regarding the goal. Hence the weight vector's entries are 1/5 for each of the criteria. It is thus obvious that the ranking of genes is automatically normalized and it shows the significance of each gene taking into account not only a single criterion but all criteria simultaneously. The top ranking genes are then selected for classification.

3. HMMs for cancer classification

The methodology proposed in this paper is diagrammed in Fig. 2 where the modified AHP and the HMM classifier play crucial roles in feature selection and classification respectively.

An HMM, which was developed by Baum and Petrie [3], is a probabilistic model in which the system is assumed to be a Markov process with hidden states. HMM is used for representing probability distribution over sequences of observations.

Table 1
Five eigenvectors of the option performance matrix.

	t-test	Entropy	ROC	Wilcoxon	SNR
Gene 1	ϵ_{T1}	ϵ_{E1}	ϵ_{R1}	ϵ_{W1}	ϵ_{S1}
...	...	...	...	...	...
Gene n	ϵ_{Tn}	ϵ_{En}	ϵ_{Rn}	ϵ_{Wn}	ϵ_{Sn}

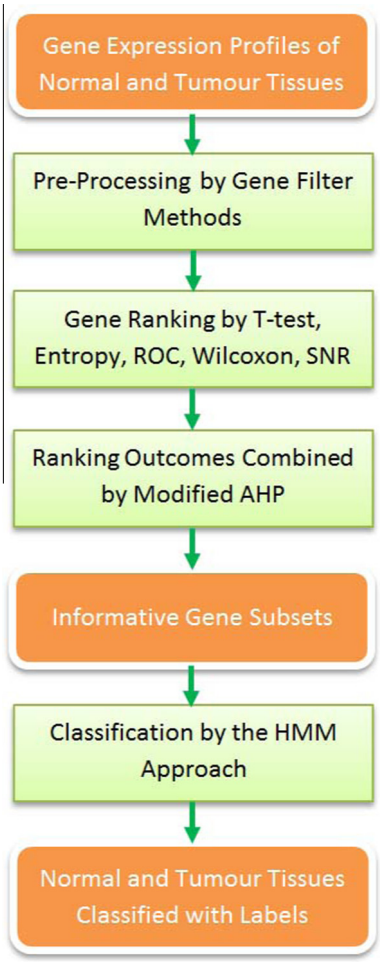

Fig. 2. Cancer classification methodology.

The following key elements characterize an HMM:

- N: the number of states of the model.
- M: the number of observation symbols in the alphabet. M is infinite if the observations are continuous.
- $\pi = \{\pi_i\}$: the initial state distribution where $\pi_i = P(q_1 = s_i),\ 1 \leqslant i \leqslant N$ is the probability of s_i being the first state of a state sequence.
- $A = \{a_{ij}\}$: the transition probability matrix whose element a_{ij} represents the probability to go from state i to sate $j : a_{ij} = P(q_{t+1} = s_j | q_t = s_i),\ 1 \leqslant i,j \leqslant N$. Transition probabilities must satisfy the normal stochastic constraints: $a_{ij} \geqslant 0$, $1 \leqslant i,j \leqslant N$ and $\sum_{j=1}^{N} a_{ij} = 1,\ 1 \leqslant i \leqslant N$.
- $B = \{b_j(k)\}$: the emission probability matrix where $b_j(k)$ specifies the likelihood of the kth observation symbol, v_k, in the alphabet if the model is in state $s_j : b_j(k) = P(O_t = v_k | q_t = s_j),\ 1 \leqslant j \leqslant N,\ 1 \leqslant k \leqslant M$ (e.g. see Fig. 3 for an HMM with 3 states and their discrete emission probabilities through the observation sequence).

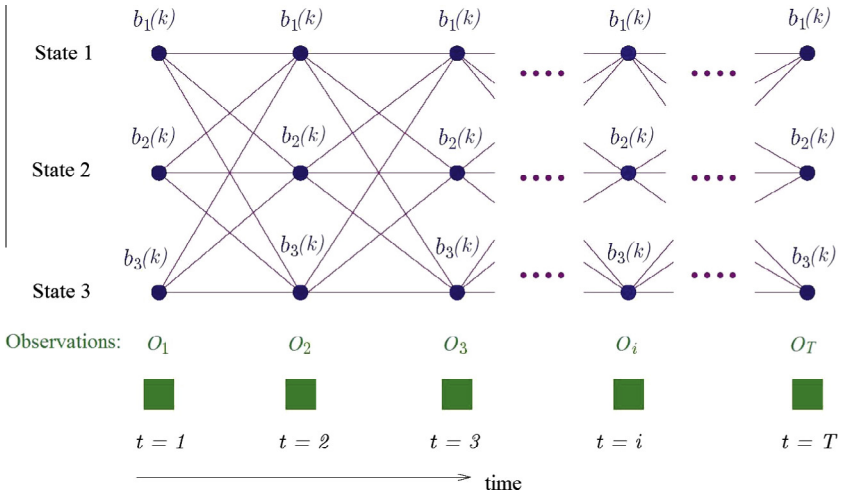

Fig. 3. Example of an HMM with three states.

Elements of B should satisfy the following stochastic constraints: $b_j(k) \geqslant 0,\ 1 \leqslant j \leqslant N,\ 1 \leqslant k \leqslant M$ and $\sum_{k=1}^{M} b_j(k) = 1,\ 1 \leqslant j \leqslant N$.

For continuous valued observations, a continuous probability density function is used instead of a set of discrete probabilities. Then parameters of the probability density function must be specified. The probability density is often approximated by a weighted sum of M Gaussian distributions $\mathcal{N}$: $b_j(O_t) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mu_{jm}, \Sigma_{jm}, O_t)$ where c_{jm} are the weighting coefficients, μ_{jm} is the mean vector and Σ_{jm} is the covariance matrix. The stochastic process induces constraints $c_{jm} \geqslant 0,\ 1 \leqslant j \leqslant N,\ 1 \leqslant m \leqslant M$ and $\sum_{m=1}^{M} c_{jm} = 1,\ 1 \leqslant j \leqslant N$.

The observation O_t at time t is generated by some process whose state s_t is hidden from the observer. The state of this hidden process satisfies the Markov chain properties. Given the present state s_t, we have that the observation O_t is independent of other states:

$$P(O_t|s_1, s_2, \ldots, s_t) = P(O_t|s_t) \tag{14}$$

and that the future state is independent of the past (i.e. a first order HMM):

$$P(s_{t+1}|s_1, s_2, \ldots, s_t) = P(s_{t+1}|s_t) \tag{15}$$

Given observation sequence $O = (O_1, O_2, \ldots, O_T)$ where T is the length of the sequence, the model $\Omega = (A, B, \pi)$ and the state sequence $S = (s_1, s_2, \ldots, s_T)$, one can compute the likelihood of the observation sequence as follows:

$$P(O|\Omega) = \sum_{S} P(O, S|\Omega) = \sum_{S} P(O|S, \Omega)P(S|\Omega) \tag{16}$$

The Markov chain properties (i.e. Eqs. (14) and (15)) allow to calculate the following propabilities as:

$$P(O|S, \Omega) = \prod_{t=1}^{T} P(O_t|s_t, \Omega) = \prod_{t=1}^{T} b_t(O_t) \tag{17}$$

$$P(S|\Omega) = \pi_1 \prod_{t=1}^{T-1} a_{t,t+1} \tag{18}$$

Therefore, Eq. (16) can be extended as follows:

$$P(O|\Omega) = \sum_{S} P(O|S, \Omega)P(S|\Omega) = \sum_{s_1 s_2 \ldots s_T} \pi_1 \prod_{t=1}^{T} a_{t,t+1} b_t(O_t) \tag{19}$$

The complexity of this calculation is $O(\mathrm{TN}^T)$, which is remarkably intensive. The likelihood $P(O|\Omega)$ may be computed by a more efficient method using forward–backward procedure based on the principle of dynamic programming algorithm [22].

Define the forward probability as the joint probability of observing the first t symbols and being in state k at time t

$$\alpha_k(t) = P(O_1, O_2, \ldots, O_t, s_t = k) \tag{20}$$

This probability can be evaluated by the following recursive formulae:

$$\alpha_k(1) = \pi_k b_k(O_1), \quad 1 \leqslant k \leqslant N \tag{21}$$

$$\alpha_k(t) = b_k(O_t)\sum_{l=1}^{N}\alpha_l(t-1)a_{l,k}, \quad 1 \leqslant t \leqslant T, \ \ 1 \leqslant k \leqslant N \tag{22}$$

Define the backward probability $\beta_k(t)$ as the conditional probability of observing the symbols after time t, given the state at time t is k

$$\beta_k(t) = P(O_{t+1}, \ldots, O_T | s_t = k), \quad 1 \leqslant t \leqslant T-1 \tag{23}$$

while $\beta_k(T) = 1$ for all k.

Similar with the forward probability, the backward probability can be calculated by the following recursion

$$\beta_k(T) = 1 \tag{24}$$

$$\beta_k(t) = \sum_{l=1}^{N} a_{k,l} b_l(O_{t+1})\beta_l(t+1), \quad 1 \leqslant t \leqslant T-1 \tag{25}$$

Then for any t, we have that

$$P(O|\Omega) = \sum_{k=1}^{N}\alpha_k(t)\beta_k(t) \tag{26}$$

In particular, if $t = T$,

$$P(O|\Omega) = \sum_{k=1}^{N}\alpha_k(T)\beta_k(T) = \sum_{k=1}^{N}\alpha_k(T) \tag{27}$$

This method requires $O(TN^2)$ computation, which is much simpler than the conventional approach by Eq. (19).

In the application of HMM for classification, for the training data of each class label, we fit a corresponding ergodic HMM using the expectation–maximization (Baum–Welch) learning algorithm [4]. The Baum–Welch algorithm employs the calculation process of forward and backward probabilities presented above to adjust parameters. The parameters are tuned to maximize the probability of the observation sequence for a given HMM model. This optimization approach is known as the maximum likelihood criterion.

Once the training process is completed, given a new testing observation, the likelihood of the observation with respect to HMMs are calculated using Eq. (27). The class label of the new observation is identified by the HMM that yields the greatest likelihood.

4. Experiments and results

4.1. Performance evaluation

It is well-known that there are often a small number of samples in gene expression microarray datasets. In order to train as many examples as possible, the leave one out cross validation (LOOCV) [16,13] is employed. The strategy divides all samples at random into K distinct subsets, where K is equal to the number of samples. As with traditional k-fold cross validation, this strategy uses $K-1$ subsets for training whilst the kth sample is for testing. The LOOCV accuracy, denoted as ACC, is calculated as follows:

$$ACC = A/K \tag{28}$$

where A is the number of correctly classified samples in K experiments.

Sensitivity and specificity are also utilized to measure performance of classification techniques. The sensitivity of a test refers to the proportion of patients with disease who test positive. Conversely, specificity refers to the proportion of patients without disease who test negative. Another important performance metric in medical application, which is area under the ROC curve (AUC) is also calculated.

For unbiased comparisons among gene selection methods and classifiers, each classifier is repeated 20 times on a gene subset and the average performance is reported. Accordingly, for each gene subset, a classifier is performed in total $20 \times K$ times to generate 20 independent results.

4.2. Datasets

Four benchmark datasets used for experiments include diffuse large B-cell lymphomas (DLBCL) [18], leukemia cancer [15], colon [1] and prostate [26].

DLBCL and follicular lymphomas (FL) are two B-cell lineage malignancies that expose different clinical presentations, natural histories and response to therapy. However, FLs frequently evolve over time and acquire the morphologic and clinical features of DLBCLs and some subsets of DLBCLs have chromosomal translocations characteristic of FLs. The gene-expression based classification models were built to distinguish between these two lymphomas. The DLBCL dataset is composed of 7070 genes and 77 samples where DLBCL contributes 75.3% with 58 samples and the rest 24.7% with 19 instances are of FLs.

The leukemia dataset contains information on gene-expression from human acute myeloid (AML) and acute lymphoblastic (ALL) patients. The dataset consists of 5147 genes with 72 samples of which 47 samples (65.3%) are ALL and the remaining are AML with 25 tissues (34.7%).

Alternatively, the prostate dataset comprises 12,533 genes with 102 samples. Among them, prostate tumor occupies 51% with 52 samples. The normal tissues account for 49% with 50 samples.

The colon dataset provides the expression levels of the 2000 genes with the highest minimal intensity across 62 tissues. Normal tissues account for 35.5% with 22 samples whilst the rest 40 tissues (64.5%) are colon cancer samples.

In the data pre-processing step, some filter approaches are employed to remove genes with low absolute values, little variation, small profile ranges or low entropy. These genes are generally not of interest because their quality is often bad due to large quantization errors or simply poor spot hybridization [17]. The gene profiles are then normalized using the quantile normalization technique [6].

4.3. Results and discussions

4.3.1. Comparisons among gene selection methods

Popular gene selectors including information gain (IG), symmetrical uncertainty (SU), ReliefF [30] and Bhattacharyya distance (BD) [10] are also implemented to compare with the proposed modified AHP. For comparisons, we select the same number of top ranked genes for every method. The five most informative genes are chose to form the gene subsets for all classifiers. Projections on the 3D space of gene subsets selected by IG, SU, BD, ReliefF and AHP on the DLBCL dataset are exhibited in Figs. 4–8 respectively. Gene expression profiles on the axes of these figures are shown after the quantile normalization.

There is a clear separation of two classes in the AHP projection (see Fig. 8). In contrast, projections of gene subsets by other methods display vague distinctions between two classes (see Figs. 4–7). As quality of features substantially affects the classification accuracy, this advantage of the modified AHP would improve classification performance of classifiers that take selected features as inputs.

Fig. 9 shows box plots demonstrating accuracy and AUC obtained when applying the HMM classifier to gene subsets selected by IG, SU, BD, ReliefF, and AHP. In all four datasets, we see a considerable dominance of the AHP against other gene selectors. BD and ReliefF attain mediocre performance as they provide results that are inferior to AHP but better than IG and SU.

Moreover, the relatively small interquartile ranges of the AHP boxes compared to those of other methods demonstrate the stability and robustness of the modified AHP method. For example, in the DLBCL, colon and prostate datasets, the boxes of AHP are the smallest ones among boxes of five gene selectors in both criteria: accuracy and AUC. In the leukemia dataset, the box of the AHP accuracy is larger than that of BD and ReliefF but these two methods generate several outliers.

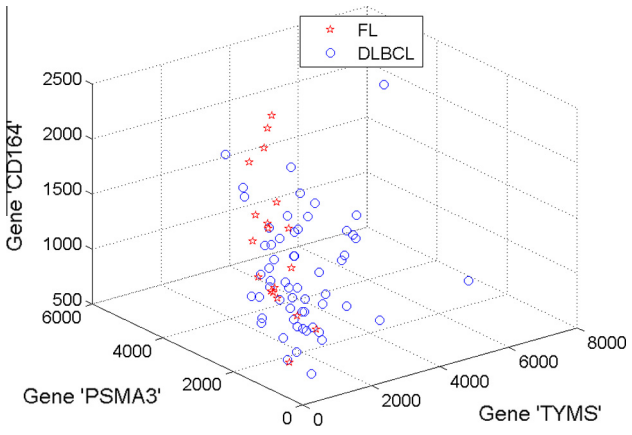

Fig. 4. IG gene 3D projection in DLBCL dataset.

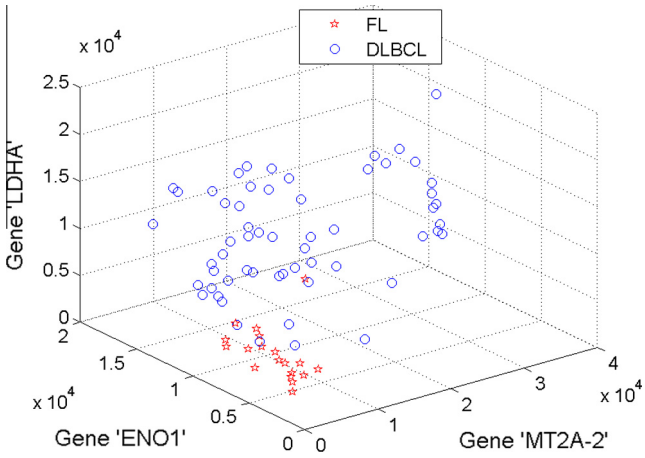

Fig. 5. SU gene 3D projection in DLBCL dataset.

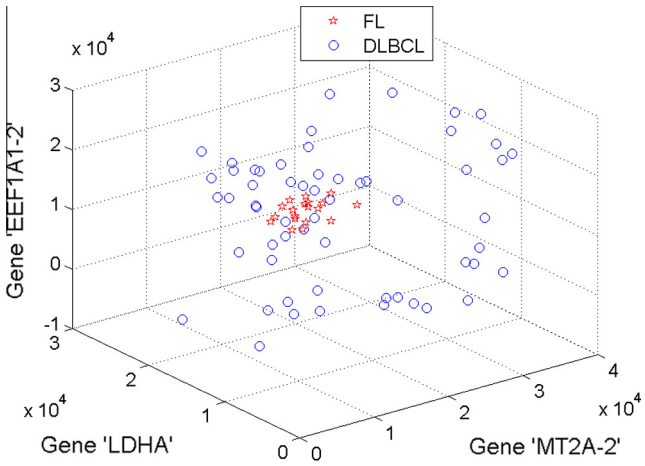

Fig. 6. BD gene 3D projection in DLBCL dataset.

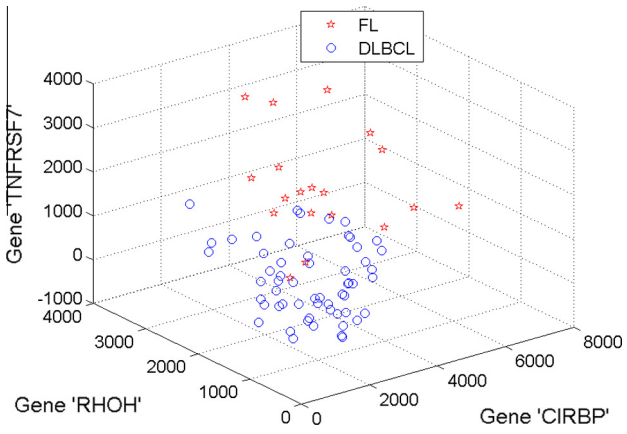

Fig. 7. ReliefF gene 3D projection in DLBCL dataset.

Processing time required by gene selectors in four datasets is shown in Fig. 10. For each method, the time required is proportional to the size, i.e. the number of genes and samples, of the dataset. BD is the fastest method as it needs just several milliseconds to accomplish the selection. AHP spends more or less 1 s depending on the dataset and it takes less time amount

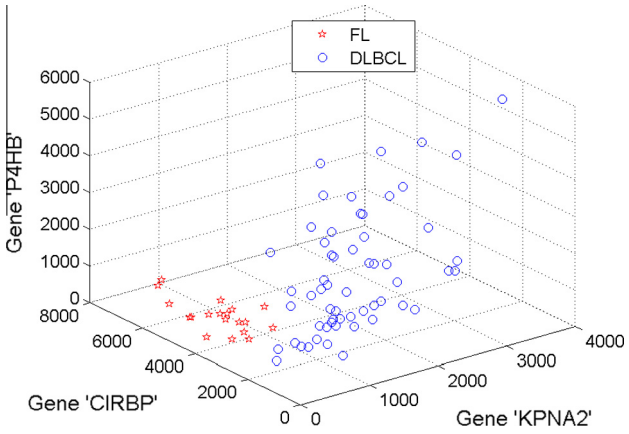

Fig. 8. AHP gene 3D projection in DLBCL dataset.

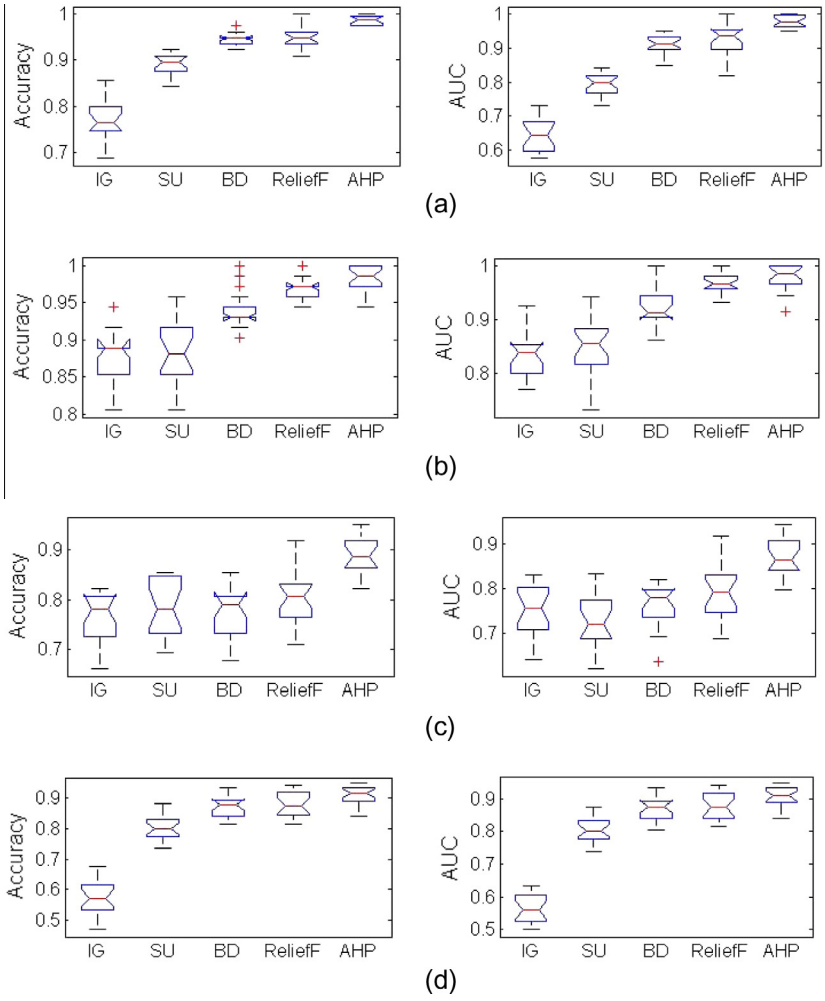

Fig. 9. Results across different gene selectors in (a) DLBCL, (b) leukemia, (c) colon, (d) prostate datasets.

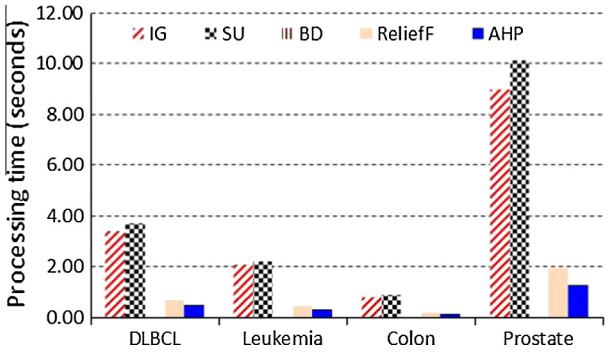

Fig. 10. Processing time of gene selectors.

compared to ReliefF in every dataset. IG and SU are the worst gene selectors in terms of time consumption as they take much longer time than other methods. Modified AHP is therefore the second best method regarding processing time.

Consequently, AHP demonstrates advantages compared to other gene selectors as it leads to greater classification performance. It also yields comparatively stable results and requires a relatively low computational cost.

4.3.2. Comparisons among classifiers

To compare with the HMM classification approach, a range of prevalent classifiers such as k-nearest neighbors (kNN) [2], probabilistic neural network (PNN) [27], support vector machine (SVM) [11], multilayer perceptron (MLP) [21], fuzzy ARTMAP (FARTMAP) [8], and ensemble learning AdaBoost [14] are implemented. The Euclidean distance metric is utilized in the implementation of the kNN classifier. Alternatively, the method used to find the separating hyperplane in SVM is the sequential minimal optimization, which implements the L^1 soft-margin SVM classifier. The box constraint parameter C for the SVM soft margin is equivalent to 1. On the other hand, the decision tree is used as a base learning algorithm of the AdaBoost ensemble learning.

For the HMM classifier, two HMMs corresponding to two class labels (cancer and normal) are constructed and trained using the training data of each dataset. After training, each testing observation is fed to these two HMMs. Class label is determined based on the HMM generating the greater likelihood of the observation.

Results of experiments on four datasets, i.e. DLBCL, leukemia, colon and prostate, are reported in Tables 2–5 respectively. Each result in these tables is the mean plus or minus the standard deviation of 20 independent outcomes of a classifier performed on a gene subset selected by the modified AHP.

HMM clearly demonstrates a superior performance in terms of accuracy and AUC compared to other classifiers. Application of HMM in the DLBCL dataset yields the accuracy and AUC at 98.83% and 98.14%, which are the maximum results in this dataset. There is often a trade-off between the sensitivity and specificity criteria. In practice, a method that provides greater sensitivity often give smaller specificity than the other. This can be found in the results presented in Tables 2–5. For example, in the DLBCL dataset (Table 2), although HMM is significantly better than SVM, the comparisons between sensitivity and specificity does not give a consistent conclusion. SVM achieves the maximum specificity at 100%, which is greater than that of HMM at 99.32%. In contrast, SVM's sensitivity is considerably inferior to that of HMM: 89.54% versus 96.97%.

In addition to the greater performance, HMM also produces relatively smaller standard deviation results. This shows the stability of the HMM classifier. In the DLBCL dataset, HMM generates one of the smallest accuracy and AUC standard deviations, at 1.33% and 2.20% respectively (Table 2). These values are also rather small in the other datasets. For example, in the leukemia dataset, HMM's accuracy standard deviation is inferior (larger) to that of kNN but it is superior (smaller) to those of PNN, SVM, MLP, FARTMAP and AdaBoost (Table 3). Likewise, HMM's accuracy standard deviation in the colon dataset is larger than those of kNN and PNN but it is smaller than those of other four methods (Table 4). Particularly, in the prostate dataset, HMM obtains the smallest AUC standard deviation at 2.20% compared with those of other classifiers (see Table 5).

Table 2
Results (in %) of the DLBCL dataset.

Classifiers	Sensitivity	Specificity	AUC	Accuracy	*p*-values
kNN	94.03 ± 5.88	98.18 ± 1.72	96.10 ± 3.17	97.01 ± 2.23	0.0057
PNN	94.93 ± 6.33	96.94 ± 2.52	95.94 ± 3.85	96.56 ± 2.84	0.0104
SVM	89.54 ± 7.95	100.0 ± 0.00	94.77 ± 3.98	97.34 ± 2.17	0.0077
MLP	94.30 ± 4.82	97.00 ± 2.45	95.65 ± 2.94	96.30 ± 2.36	0.0003
FARTMAP	93.63 ± 4.51	100.0 ± 0.00	96.81 ± 2.26	98.44 ± 1.08	0.1676
AdaBoost	95.20 ± 5.14	98.81 ± 1.45	97.00 ± 2.61	97.79 ± 1.53	0.0135
HMM	96.97 ± 4.07	99.32 ± 1.04	98.14 ± 2.20	98.83 ± 1.33	–

Table 3
Results (in %) of the leukemia dataset.

Classifiers	Sensitivity	Specificity	AUC	Accuracy	*p*-values
kNN	96.39 ± 3.41	97.76 ± 1.81	97.07 ± 1.85	97.36 ± 1.62	0.0608
PNN	93.75 ± 4.80	98.04 ± 1.75	95.90 ± 2.36	96.53 ± 1.94	0.0049
SVM	77.25 ± 6.25	100.0 ± 0.00	88.62 ± 3.13	92.78 ± 2.19	0.0000
MLP	90.70 ± 5.75	96.79 ± 2.93	93.74 ± 3.17	94.72 ± 2.80	0.0000
FARTMAP	92.15 ± 5.52	98.17 ± 2.36	95.16 ± 2.99	95.97 ± 2.54	0.0002
AdaBoost	100.0 ± 0.00	95.08 ± 3.40	97.54 ± 1.70	96.74 ± 2.31	0.0146
HMM	96.48 ± 4.43	99.13 ± 1.09	97.81 ± 2.34	98.26 ± 1.68	–

Table 4
Results (in %) of the colon dataset.

Classifiers	Sensitivity	Specificity	AUC	Accuracy	*p*-values
kNN	78.27 ± 8.16	90.79 ± 4.23	84.53 ± 4.67	86.53 ± 3.89	0.0639
PNN	73.16 ± 9.78	94.13 ± 3.55	83.64 ± 4.65	86.45 ± 3.45	0.0384
SVM	81.18 ± 8.67	84.46 ± 5.63	82.82 ± 5.77	83.47 ± 5.23	0.0018
MLP	66.29 ± 8.80	85.11 ± 4.87	75.70 ± 4.66	78.55 ± 4.53	0.0000
FARTMAP	68.85 ± 10.15	84.67 ± 6.80	76.76 ± 6.44	78.95 ± 6.11	0.0000
AdaBoost	70.69 ± 11.32	84.71 ± 6.00	77.70 ± 6.94	80.24 ± 5.89	0.0000
HMM	81.47 ± 8.58	92.83 ± 4.27	87.15 ± 5.34	89.11 ± 4.47	–

Table 5
Results (in %) of the prostate dataset.

Classifiers	Sensitivity	Specificity	AUC	Accuracy	*p*-values
kNN	87.79 ± 4.58	90.55 ± 3.76	89.17 ± 3.22	89.26 ± 3.14	0.0034
PNN	82.96 ± 5.25	95.08 ± 2.73	89.02 ± 2.99	89.22 ± 2.84	0.0029
SVM	95.10 ± 2.71	82.40 ± 5.05	88.75 ± 2.99	88.77 ± 2.92	0.0007
MLP	91.82 ± 4.76	89.07 ± 3.69	90.45 ± 3.25	90.34 ± 3.25	0.0508
FARTMAP	82.83 ± 5.08	95.49 ± 2.36	89.16 ± 2.52	89.36 ± 2.76	0.0033
AdaBoost	87.14 ± 5.40	93.28 ± 2.90	90.21 ± 3.11	90.25 ± 3.24	0.0410
HMM	95.60 ± 2.59	88.84 ± 4.27	92.22 ± 2.20	92.01 ± 2.59	–

To derive convincing conclusions regarding the dominance of HMM compared to other classifiers, we implement the Kruskal–Wallis test for comparing two sets of accuracy results. The Kruskal–Wallis test is a nonparametric version of the classical one-way ANOVA. As the results over 20 trials may not be normally distributed, they may violate the normal assumption of the ANOVA. Therefore the use of Kruskal–Wallis test is more appropriate. The test returns the *p*-value for the null hypothesis that all samples in two sets of results are drawn from the same population (or from different populations with the same mean). The *p*-values of the Kruskal–Wallis test are reported in the last column of Tables 2–5. Note that the test is performed to compare the set of 20 outcomes generated by HMM against that obtained by each of the competing classifiers. For example, in the DLBCL dataset (Table 2), the *p*-value in the kNN row (0.0057) represents the result of the Kruskal–Wallis test performed on two sets (populations) of classification results: one set comprises 20 outcomes of HMM and another set consists of 20 outcomes of kNN. More specifically, this *p*-value shows that there is a significant difference (at 5% level) between the two populations. It also means that HMM statistically dominates kNN at the 5% significance level.

Through all experiments, there is only one out of six cases where the Kruskal–Wallis test results do not reject the null hypothesis. Other than that, most of the tests reject the null hypothesis that results of two methods (HMM and each of the competing classifiers) come from the same distribution at the 5% significance level. Fig. 11 graphically details the performance comparisons among HMM and other competing classifiers. In each dataset, there are two plots representing results in terms of accuracy and AUC criteria. Each box in these plots shows the median and distribution of results throughout 20 trials.

The median values of the HMM boxes are greater than those of other methods in all four datasets. This is consistent with the mean values reported in Tables 2–5 that show a significant dominance of HMM against other classifiers.

Fig. 12 shows processing time by classifiers in different datasets. Colon is the smallest dataset so that it requires the smallest time quantity by classifiers. In contrast, classifiers spend the largest amount of time for the prostate dataset as it has the largest size. Classification methods such as kNN, PNN, SVM and FARTMAP can process whole datasets in a few seconds. AdaBoost is the most intensive computation method as it is an ensemble learning method. For instance, in the DLBCL dataset, it needs more than 55 s to complete classification of the whole dataset. The maximum time amount required is for processing the prostate dataset, which is more than 90 s. MLP is the second most computationally expensive method. It takes more than 10 s for the DLBCL dataset and more than 25 s for the prostate dataset. HMM requires relatively low

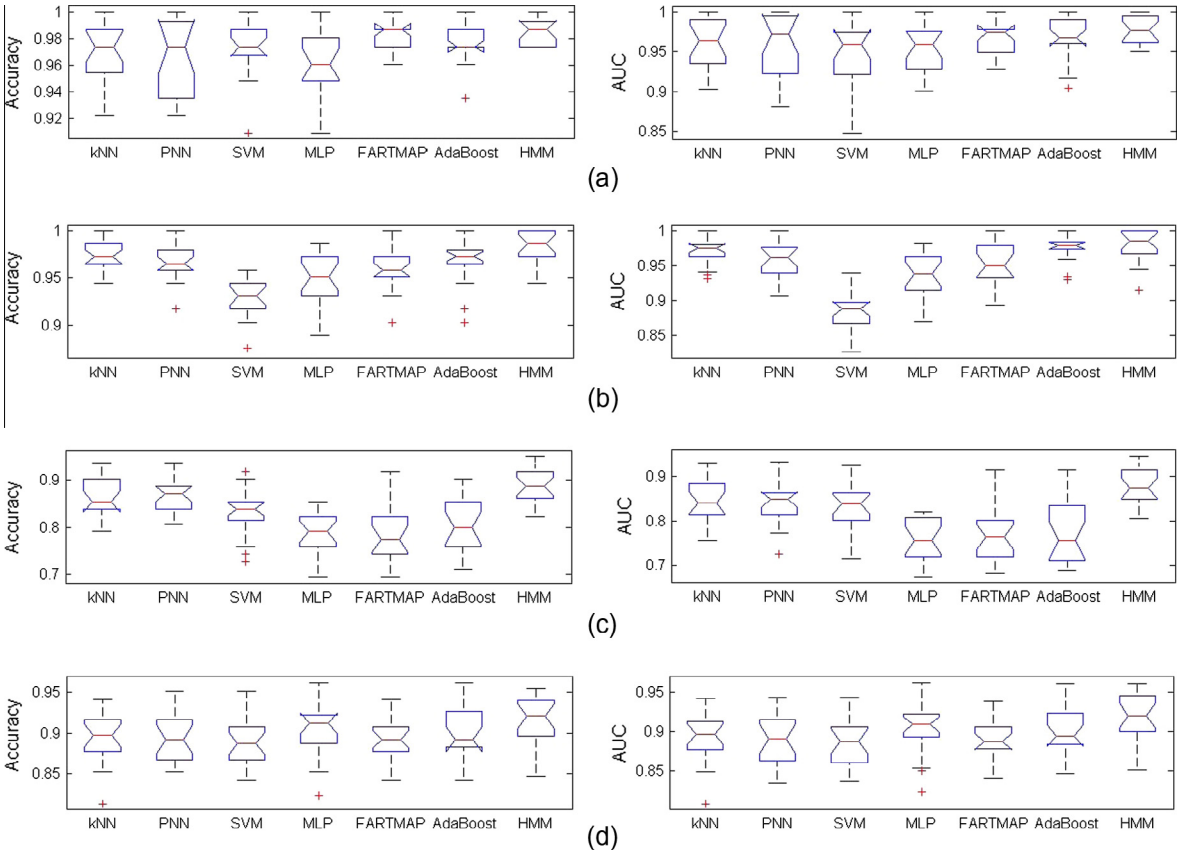

Fig. 11. Box plot comparisons among classifiers in (a) DLBCL, (b) leukemia, (c) colon, (d) prostate datasets.

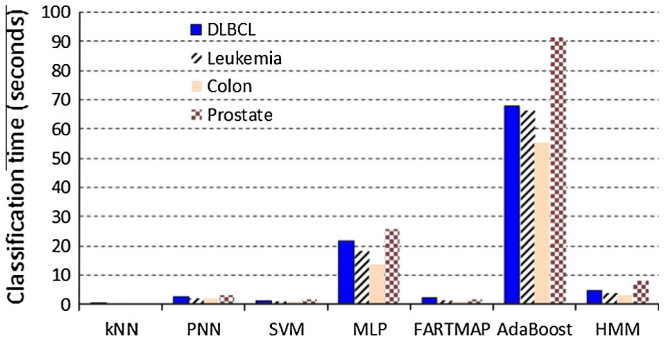

Fig. 12. Processing time of classifiers.

computational costs. It needs several seconds to train models and classify all samples of datasets. Although HMM takes more time than kNN, PNN, SVM and FARTMAP, it significantly dominates MLP and AdaBoost regarding computational expense.

5. Conclusions

With an objective quantitative ranking procedure, modified AHP provides more rigorous and unbiased assessments in building the reciprocal comparison matrix without the need for expert knowledge. It is an aggregation method that incorporates advantages and quintessence of every individual method. Competing gene selectors IG, SU, BD and ReliefF are dominated by AHP in terms of classification performance. Gene subsets selected by AHP are robust and stable and play an important role in improving the performance of classifiers applied subsequently.

Amongst seven investigated classifiers, HMM proves to be the most robust method. It yields greater accuracy and AUC and also relatively stable results compared to other methods. Noticeably, application of HMM takes less time amount than that of MLP and AdaBoost although it provides greater performance than these two methods. Other classifiers such as kNN, PNN, SVM, and FARTMAP are less time-consuming than HMM but their performance is significantly inferior to that of HMM. The design of HMM for classification meets three fundamental objectives of a classifier: providing great classification performance but not sacrificing result stability and at the same time restraining the computational costs.

The improved performance of AHP-HMM is achieved by the stability and robustness of the combination between modified AHP-based gene selection and the HMM classifier. This combination strengthens the cancer classification performance by individually improving not only the efficiency of the gene selection (i.e. modified AHP) but also that of the classifier (i.e. HMM). The proposed approach therefore can be implemented in the real clinical practice as a useful software tool for early detection, understanding, and treatment of cancers in an effective and efficient manner. It in general will contribute to improve the public health by increasing human longevity, reducing mortality rate in communities and ensuring people live healthier and more independent lives.

Acknowledgments

This research is supported by the Australian Research Council (Discovery Grant DP120102112) and the Centre for Intelligent Systems Research (CISR) at Deakin University.

References

[1] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Natl. Acad. Sci. 96 (12) (1999) 6745–6750.
[2] N.S. Altman, An introduction to kernel and nearest-neighbor nonparametric regression, Am. Stat. 46 (3) (1992) 175–185.
[3] L.E. Baum, T. Petrie, Statistical inference for probabilistic functions of finite state Markov chains, Ann. Math. Stat. 37 (6) (1966) 1554–1563.
[4] 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 (1) (1970) 164–171.
[5] V. Bolon-Canedo, N. Snchez-Maroo, A. Alonso-Betanzos, J.M. Bentez, F. Herrera, A review of microarray datasets and applied feature selection methods, Inf. Sci. 282 (2014) 111–135.
[6] B.M. Bolstad, R.A. Irizarry, M. strand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics 19 (2) (2003) 185–193.
[7] F.J. Burguillo, L.A. Corchete, J. Martin, I. Barrera, W.G. Bardsley, A partial least squares algorithm for microarray data analysis using the vip statistic for gene selection and binary classification, Curr. Bioinform. 9 (3) (2014) 348–359.
[8] G.A. Carpenter, S. Grossberg, N. Markuzon, J.H. Reynolds, D.B. Rosen, Fuzzy ARTMAP: a neural network architecture for incremental supervised learning of analog multidimensional maps, IEEE Trans. Neural Netw. 3 (5) (1992) 698–713.
[9] K.H. Chen, K.J. Wang, M.L. Tsai, K.M. Wang, A.M. Adrian, W.C. Cheng, T.S. Yang, N.C. Teng, K.P. Tan, K.S. Chang, Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm, BMC Bioinform. 15 (49) (2014) 1–10.
[10] E. Choi, C. Lee, Feature extraction based on the Bhattacharyya distance, Pattern Recogn. 36 (8) (2003) 1703–1709.
[11] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[12] L. Deng, J. Pei, J. Ma, D.L. Lee, A rank sum test method for informative gene discovery, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, vol. 1, 2004, pp. 410–419.
[13] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinform. Comput. Biol. 3 (2) (2005) 185–205.
[14] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.
[15] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (5439) (1999) 531–537.
[16] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nature Med. 7 (6) (2001) 673–679.
[17] I. Kohane, A. Kho, A. Butte, Microarrays for an Integrative Genomics, first ed., MIT Press, Cambridge, MA, 2003.
[18] S. Monti, K.J. Savage, J.K. Kutok, Molecular profiling of diffuse large b cell lymphoma reveals a novel disease subtype with brisk host inflammatory response and distinct genetic features, Blood 105 (5) (2005) 1851–1861.
[19] D.V. Nguyen, D.M. Rocke, Multi-class cancer classification via partial least squares with gene expression profiles, Bioinformatics 18 (9) (2002) 1216–1226.
[20] T.T. Nguyen, L. Gordon-Brown, Constrained fuzzy hierarchical analysis for portfolio selection under higher moments, IEEE Trans. Fuzzy Syst. 20 (4) (2012) 666–682.
[21] S.K. Pal, S. Mitra, Multilayer perceptron, fuzzy sets, and classification, IEEE Trans. Neural Netw. 3 (5) (1992) 683–697.
[22] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.
[23] J.C. Rajapakse, P.A. Mundra, Multiclass gene selection using pareto-fronts, IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB) 10 (1) (2013) 87–97.
[24] T.L. Saaty, K. Peniwati, Group Decision Making: Drawing Out and Reconciling Differences, first ed., RWS Publications, Pittsburgh, PA, 2008.
[25] S.S. Shreem, S. Abdullah, M.Z.A. Nazri, Hybridising harmony search with a Markov blanket for gene selection problems, Inf. Sci. 258 (2014) 108–121.
[26] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, Gene expression correlates of clinical prostate cancer behavior, Cancer Cell 1 (2) (2002) 203–209.
[27] D.F. Specht, Probabilistic neural networks, Neural Netw. 3 (1) (1990) 109–118.
[28] S. Sun, Q. Peng, A. Shakoor, A kernel-based multivariate feature selection method for microarray data classification, PloS One 9 (7) (2014) e102541.
[29] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Academic Press, San Diego, USA, 1999.
[30] Y. Wang, I.V. Tetko, M.A. Hall, E. Frank, A. Facius, K.F. Mayer, H.W. Mewes, Gene selection from microarray data for cancer classification-a machine learning approach, Comput. Biol. Chem. 29 (1) (2005) 37–46.
[31] K.Y. Yeung, R.E. Bumgarner, A.E. Raftery, Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data, Bioinformatics 21 (10) (2005) 2394–2402.
[32] W. You, Z. Yang, M. Yuan, G. Ji, Totalpls: local dimension reduction for multicategory microarray data, IEEE Trans. Hum.–Mach. Syst. 44 (1) (2014) 125–138.