IDENTIFYING PREDICTIVE GENES FOR SEQUENCE CLASSIFICATION USING ARTIFICIAL IMMUNE RECOGNITION SYSTEM

Canan BATUR
Yıldız Technical University, Department of Computer Engineering, Istanbul- TURKEY
canan@ce.yildiz.edu.tr

Banu DİRİ
Yıldız Technical University, Department of Computer Engineering, Istanbul- TURKEY
banu@ce.yildiz.edu.tr

Abstract: The small sample data in the high-dimensional data space are encountered in biological applications such as in gene expression microarrays and proteomics mass spectrometry. Due to the fact that such data have characteristics such as high-dimensionality and small sample dimension, their classification becomes hard. Many feature selection algorithms were developed for the purpose of reducing the dimensionality of this kind of data and improving the accuracy of classifiers. In the realization of area discoveries through feature sets, the selected feature subsets skip important information in unnecessary feature sets. This problem comes into prominence with the feature, in the process of performing the discovery of information from the high-dimensional data space. This paper evaluates the proposed ensemble gene selection method based on a local feature selection to Artificial Immune Recognition algorithms in order to find the optimal biological sequences. The unique feature of this study is developing the different type of associated feature groups defined using high-dimensional data in order to find important tumor-related genes. The comparative tests were performed on the training set and test set separately with using support vector machines and k-NN classifiers.

Keywords: Group Based Learning, Gene Selection, Sequence Classification, Machine Learning

Introduction

The discovery of biomarkers from high-dimensional data is an important research topic in the biomedical field. Selecting the most distinctive or critical features for classification is a feature selection problem. Many feature selection algorithms were developed for the purpose of reducing the dimensionality of this kind of data and improving the accuracy performance of the classifiers. A gene selection framework must be established in order to select the most discriminating genes in order to identify biomarkers.

Feature selection is a problem of minimum subset selection from the original feature set for the best accuracy estimation. Generally high dimensional feature selection methods can be classified into two categories: Filters and Wrappers. For filter methods, evaluation of the feature discriminability depends on only the inherent features of the microarray data and subclass information, such as density, correlation, Chi-square statistics, and relief. Wrapper methods search for an optimal feature subset by the evaluation function of a learning algorithm for assessing the goodness of a feature subset. Filters sometimes used as a pre-processing step for other approaches and usually fast.

The idea of the ensemble gene selection framework is converging to original feature groups by creating a set of feature groups is based on the principles of group-based learning method. Ensemble gene selection method depends on feature group formation and feature selection procedures in order to improve the model. In feature grouping stage, different samplings of the original data are generated to create different subset groups. Identifying different type of feature groups relying on data-driven or knowledge-driven group formation method. The data-driven group formation method, exploits the characteristics of target data and the knowledge-driven group formation method is to find group of associated genes or proteins that have coherent expression pattern in the same pathway (He, Yu, 2010). Within the scope of this work, data-driven feature group formation used to pre-select the different type associated gene subset groups. Each type of associative feature group is improved by being optimized with Artificial Immune Recognition Systems and learning is performed at the group level. Within the scope of this study, in the feature selection framework the feature groups were taken as a basis. An attempt to develop the associated feature groups defined using high-dimensional data based on a local feature selection to Artificial Immune Recognition System (LFSAIRS1, LFSAIRS2, Parallel-LFSAIRS1, Parallel -LFSAIRS2).

Creating different type associated feature groups from high-dimensional data is mentioned as a associative feature groups in the second part of the paper, the Artificial immune recognition systems with proposed ensemble gene selection framework is mentioned in the third part, the data set is mentioned in the fourth part, and the comparative performance measurements of the optimal biological sequences are mentioned in the fifth part.

Associative Feature Groups

The idea of converging to original feature groups by creating a set of feature groups is based on the principles of group-based learning method. Relational features have a very high correlation in high-dimensional data sets makes it possible to use feature groups by being taken as a basis.

In this paper, the first type of the associative feature groups was obtained by the DGF (Dense Group Finder) algorithm. The main part of the DGF is the kernel density estimation and iterative mean shift procedure for all features. Density-based feature groups were obtained using the kernel density estimation designated in equation (1). The *h* parameter used for the kernel bandwidth refers to the number of nearest neighbors, *p* refers to the total number of features in the data set, *fi* refers to any feature, and *K* refers to the kernel function. *(C_j+1)* was calculated to determine the order of sequential locations of the kernel function. *(C_j+1)* positions the average by shifting it to a denser peak point using the other features in the local region determined with *h* parameter starting from the average of a certain *fi* feature (Loscalzo et al., 2009).

$$\mathrm{Cj}+1=\frac{\sum_{\mathrm{i=1}}^{\mathrm{p}}\mathrm{fi}\ \ \mathrm{K}\ \ \left(\frac{\mathrm{Cj-fi}}{\mathrm{h}}\right)}{\sum_{\mathrm{i=1}}^{\mathrm{p}}\ \ \mathrm{K}\ \ \left(\frac{\mathrm{Cj-fi}}{\mathrm{h}}\right)},\ \mathrm{j}=1,2,\ldots \qquad (1)$$

The second type of the associative feature groups was obtained by the CFG (Correlation-Based Feature Group) algorithm. The CFG is a filter-based feature selection method that sorts the feature subset by the correlation-based intuitive function. The CFG algorithm examines the usefulness of subset of attributes based on a heuristic evaluation function. In choosing a correlation-based feature, each attribute is taken into account in the correlation between the attributes, as well as the predictive predicting of the class label. The value of the heuristic evaluation function used in the evaluation of the attributes is determined by equation 2. The intuitive usability of a subset of *S* attributes with *k* attributes is represented by $meritS$, the mean attribute-class correlation is presented by *rcf* for (*f* ϵ *S*), and the correlation between the mean attributes is presented by *rff* parameters.

$$meritS=\frac{\mathrm{k*rcf}}{\sqrt{\mathrm{k+(k-1)*rff}}} \qquad (2)$$

The third type of the associative feature groups was obtained by the IGFG (Information Gain-Based Feature Group) algorithm. This method selects attributes with entropy based scores. The Entropy criterion makes a choice with the help of knowledge in the feature. For this reason, the feature is treated as a distribution and its entropy is found. The entropy of the f_t attribute with *M* data can be found by equation 3.

$$E=\ \ -\sum_{i=1}^{M}(ft(i)\log(ft(i)) \qquad (3)$$

Within the scope of this study, an attempt to create feature group sets was made by the DGF, CFG and IGFG algorithms. These feature groups obtained based on the ensemble feature selection framework using data perturbation. These feature group sets were developed with the meta-dynamics of the Artificial Immune Recognition Systems like a single cell in order to find the optimal biological sequences.

Artificial Immune Recognition System

Artificial Immune Systems are a class of adaptive computer algorithm based on metaphor of the mammalian immune system. Application areas of the Artificial Immune Systems are pattern recognition, fault and anomaly detection, data mining, classification, robotics, optimization and anomaly detection. Artificial immune recognition algorithm is one version of the Artificial Immune System which is specifically designed for the classification problems.

Artificial Immune Recognition Systems (AIRS) consist of the stages of initialization, memory cell recognition, resource competition and the selection of memory cells. In this approach, an antigen represents a single data instance and allocated to the closest matching ARB (antibody) in the pool of ARBs. At the initialization stage, the data set is normalized to the range of [0,1]. After normalization, the affinity threshold is calculated by equation (4). At the next stage, antigens are presented to the storage pool with antigen training. At the memory cell recognition stage, a stimulation value is assigned to these cells by stimulating the recognition cells in the memory pool. Affinity is calculated by equation (5), the stimulation values are calculated by equation (6) and (7).

The recognition cell with the highest stimulation value is calculated by equation (8) then $\mathrm{M}_{\mathrm{cmatch}}$ cell is cloned and mutated. The number of clones is calculated by equation (9),

$$affinity\ \ threshold = \sum_{i=1}^{n} \sum_{j=j+1}^{n} \left(\frac{affinity(\mathrm{agi,agj})}{n(n+1)/_{2}} \right) \quad (4)$$

$$affinity(\mathrm{agi,agj}) = 1 - \mathrm{Euclidean}\ \ \mathrm{distance}(\mathrm{agi,agj}) \quad (5)$$

$$stim\mathrm{ulation} = 1 - \mathrm{affinity} \quad (6)$$

$$stimulation(\mathrm{mc,ag}) = \begin{cases} \mathrm{affinity}(\mathrm{mc,ag}) & \text{if } \ \mathrm{mc.class} = \mathrm{ag.class} \\ 1 - \mathrm{affinity} & \text{otherwise} \end{cases} \quad (7)$$

$$\boldsymbol{Mcmatch} = \mathbf{argmax}(\mathbf{stimulation}(\mathbf{mc,ag})) \quad (8)$$

$$\boldsymbol{numClones} = \mathbf{stimulation} * \mathbf{clonalRate} \quad (9)$$

At the resource competition stage, when mutated clones are added to the ARB (artificial recognition spheres, antibody) pool, competition begins for the time source. According to the stimulation value, limited resource assignment to the ARB pool is made according to the stimulation value. ARBs without enough resources are removed from the system. When the stop criterion is achieved, the process ends, and the ARB with the highest stimulation value is selected as the candidate memory cell. At the selection of memory cells stage dynamically and evolving developed Memory cell pool in the algorithm is used for the classification process (Brownlee, 2005).

The basic steps of the AIRS1 algorithm, the first version of artificial immune recognition systems, and the AIRS2 algorithm, the second version, are same. The main difference between them is that the ARB pool is used as a permanent resource in the AIRS1 algorithm; it is used as a temporary resource in the AIRS2 algorithm. In the case of being used as a permanent resource, ARBs remaining from previous steps cause the algorithm to spend more time by being involved in the competition for limited resources. Therefore, the complexity of the AIRS2 algorithm is less. While AIRS1 uses the mutation parameter that can be defined by the user, AIRS2 uses the concept of somatic hyper mutation where the mutation ratio of a clone is proportional to the affinity (Wang, Chen, Adrian, 2014). While the classes of clones may change after the mutation process in the AIRS1 algorithm, classes are not allowed to change in the AIRS2 algorithm. Parallel AIRS is work into exploiting the parallelism, which caused some loss in the data reduction benefits of artificial immune recognition system. The versions of the Parallel AIRS are Parallel AIRS 1 and Parallel AIRS2. Which are modeled based on the distributed nature and parallel processing feature of the mammalian immune system (Brownlee, 2005). At first, each part of the training data set is assigned to np number of processes. Thus, it is ensured that np number of the memory pool is created by running the AIRS algorithm on each process. As a result, the memory pools obtained are merged.

A. The proposed ensemble gene selection Framework

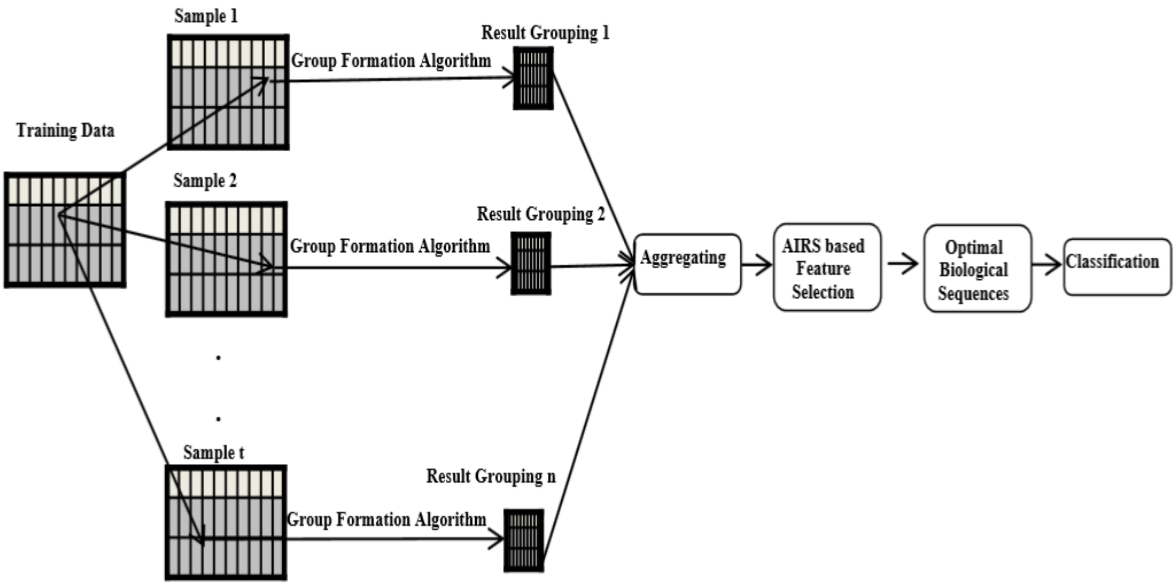
Sample 1
Result Grouping 1
Group Formation Algorithm
Training Data
Sample 2
Result Grouping 2
Group Formation Algorithm
Aggregating
AIRS based Feature Selection
Optimal Biological Sequences
Classification
Sample t
Result Grouping n
Group Formation Algorithm

Figure 1. Proposed ensemble gene selection framework

B. Standard Artificial Immune Recognition Systems:

```
Input:   InputPatterns, clone rate, mutationrate, stimthresh, resourcemax, affinitythresh
    Output: cell memor   ←eInitializeMemoryPool(InputPatterns)
         For (InputPatterni ∈ InputPatterns)
             Stimulate (cellsmemory, InputPatterns)
  cellbest esGetMostStimulatedby1NN (InputPatterni, cellsmemory)
       If cellclass,best ≠eInputPatternclass,i
           Then cellmemory   moCreateNewMemoryCell (InputPatterni )
    Else   clonenum   ← cellsstim, best x clonerate x mutationrate
             cellsclones ←lcellsbest
              For (i to clonenum )
     cellsclone ←lCloneAndMutate (cellsbest)
      End
       While (AverageStimulation (cellsclones ) ≤, stimthresh
                  For (celli ∈ cellsclones)
                   cellsclones loCloneAndMutate (cellsi)
                  End
               Stimulate (cellsclones, InputPatterns)
               ReducePoolToMaximumResources (cellsclones, resourcemax)
      End
       cellsc elGetMostStimulated (InputPatterni, cellsclones)
       If (cellsstim , c > cellsstim , best)
       Then
         cellsmemory ←ecellsc
             If (Affinity(cellsc , cell best) ≤ affinitythresh
             Then DeleteCell (cellsbest , cell memory)
             End
         End
           End
         End
         Return (cell memory)
```

The Pseudo code of the standard AIRS is represented (Wang et al, 2014).

C. *Local Feature Selection to Artificial Immune Recognition Systems: LFSAIRS*

1. The initial set of feature group sets are created based on associative group formation algorithm.
2. Do for each Antigen (Ag) until training process is completed:
    - 2.1. Calculating fitness value of the each feature set is calculated by taking into account only best matching cell
    - 2.2. Until termination do :
    - 2.3. The highest fitness value of the feature set is selected as a best feature set
    - 2.4. Generation of l clones of the best feature set
    - 2.5. Mutation of the each clone
    - 2.6. Calculating fitness of the each clone by taking into account only best candidate cell
    - 2.7. Set the highest fitness value of the feature set as a candidate optimum feature subset
    - 2.8. If best candidate cell is sufficient calculate the optimum subset then go step 3. else go 2.3
3. After memory cell replacement stage, set the optimum subset of attributes as the subset of the new attribute. If training process is completed go to Step 4, else go to step 2.
4. Selection of the best optimized feature set
5. Classification of the best optimized feature sets based on test set

Data Sets

The most common six microarray data sets were used in this study. Table 1 includes information on the genes, samples and class numbers contained in the data sets used in this study (Loscalzo et al, 2008).

Table 1: Microarray Data set

Dataset	Gene	Sample	Class
Colon	2000	62	2
Lungstd	5000	181	2
Prostate	6034	102	2
SRBCT	2308	63	4
Lymphoma	4026	62	3
Leukemia	7129	72	2

The goal of this work is identifying predictive genes for sequence classification at the group level using Artifical Immune Recognition Systems. Experimentally obtained performance values were achieved by dividing the data sets as 70% training and 30% test set. The proposed ensemble gene selection framework is applied only on training set. The goal of the designed framework is to select optimal biological sequences only on the training set in order to avoid over-fitting and selection bias problems. Therefore, the test set independent of the gene selection process. Within the scope of this work, *t* is the number of bootstraps. Bootstrap was applied on the training data set in order to ensure the resistance of training samples against variations. Then *n* is a number of the selected feature groups, which are created by DFG, CFG and IGFG algorithms. These group feature selection algorithms are separately running on the each of the bootstrap data sets. We set the *t* and *n* parameters respectively to 10 and 10 for all algorithms within the scope this paper. The number of features contained in the feature groups obtained at the end of the each group feature selection algorithm varies to for each data set. Learning was performed at the group level by improving the feature groups presented to the LFSAIRS1, LFSAIRS2, Parallel-LFSAIRS1 and Parallel-LFSAIRS2 like a single cell. Selecting an informative gene subset obtained by classification ability of a gene subset. The fitness function of each candidate solution was calculated according to the KNN classifier accuracy performance. WEKA was used to obtain classifying accuracies. For all algorithms, the classifying accuracy of the optimal candidate solution obtained at the end of each run with using the test data set with 10 cross-fold validations. The performance values added to the results were calculated by taking the average of the number of runs. Moreover, to assessment of the discrimination of more important genes is measured by its occurrence frequency of gen subset formation.

Each of the feature groups represents a candidate solution and the presence of the related feature in a feature group was encoded with 1 while the absence of it was encoded with 0. In this study, the affinity threshold value, clonal ratio, mutation ratio, np, total source, stimulating value, hypermutation ratio, run number and iteration number parameters of the Artificial Immune Recognition Systems took the values of 0.1, 10, 0.15,2, 150, 0.9, 2.0, 30 and 50, respectively.

Performance Results and Discussions

Within the scope of this study, it was focused on the problem of identifying predictive genes for sequence classification in order to find important tumor-related genes.

Table 2: Representative gene subsets obtained by LFSAIRS1 and LFSPAIRS1 based on Associative feature groups

			LFSAIRS1	**SVM**		**KNN**		**LFS PAIRS1**	**SVM**		**KNN**	
Data Set	**Group Type**	**No**	**Optimal Gene Subsets**	**10 CV % on**		**10 CV % on**		**Optimal Gene Subsets**	**10 CV % on**		**10 CV % on**	
				Train Set	**Test Set**	**Train Set**	**Test Set**		**Train Set**	**Test Set**	**Train Set**	**Test Set**
Colon	**DFG**	1	**{T63133,T57630}**	63.2	69.2	**73.4**	**92.3**	{T47377,H55759,D15057, H77536}	67.3	69.2	73.4	61.5
		2	{T53889,M22632,R83349,U10117,X14830}	61.2	69.2	73.4	61.5	{M83751, M27903}	63.2	69.2	59.1	53.8
	CFG	1	{L35545,H79349}	63.2	61.5	57.1	69.2	{U25265,T47584,X02750}	63.2	69.2	63.2	64.6
		2	{T52003,L35249,L11370}	63.2	61.5	63.2	61.5	**{H22579,X59871}**	**73.4**	**76.9**	**79.1**	**77.9**
	IGFG	1	{H55759,M83254,D15057}	63.2	69.2	69.1	76.9	{R38513,H29293,T67905, U18934}	63.2	69.2	65.3	61.5
		2	{T55558,R41561, M87434}	71.4	69.2	65.3	69.2	{R60217, J00277}	63.2	69.2	67.8	84.6
Lungstd	**DFG**	1	{32952_AT, 32980_F_AT}	86.8	88.8	95.2	88.2	**{35052_R_AT, 35053_AT, 35064_AT}**	**98.6**	**92.2**	**96.5**	**97.2**
		2	**{33052_AT, 33078_AT, 33087_S_AT}**	**95.8**	**88.8**	**96.5**	**92.8**	{35922_AT, 35932_AT, 36232_AT}	97.2	91.6	96.5	88.8
	CFG	1	{33270_I_AT, 33306_AT}	82	86.1	74.4	77.7	{32704_AT, 33270_I_AT, 33306_AT}	82	86.1	78.6	86.1
		2	{40401_AT, 40686_AT, 40700_AT}	82	88.8	80	91.6	{36370_AT, 36411_S_AT}	82	86.1	80.6	88.8
	IGFG	1	{33301_G_AT, 33307_AT, 33304_AT}	82	86.1	78.6	91.6	{AFFX-THRX-5_AT, 31439_F_AT, 31447_AT, 31477_AT}	84.8	86.1	82	88.8
		2	{32702_AT, 32711_G_AT, 32716_AT}	82	86.1	78.6	92.4	{36720_AT, 36806_AT}	82	86.1	76.5	77.7
Prostate	**DFG**	1	{32950_AT, 33469_R_AT}	80.2	71.4	71.6	71.4	{33973_AT, 34020_AT}	74	57.6	69.1	57.1
	CFG	1	{32376_AT, 32884_AT}	51.8	66.6	53.0	61.9	{41012_R_AT, 41017_AT, 41023_AT}	56.7	55	59.5	52.8
	IGFG	1	{38933_AT, 38953_AT}	49.3	47.6	76.5	71.4	{33958_AT, 33959_AT}	61.7	57.6	59.2	52.3
		2	{34930_AT, 34931_AT, 34934_AT}	42.8	46.9	60.4	71.4	{34930_AT, 34932_AT, 34934_AT 37469_AT, 37470_AT, 37471_AT}	70.3	57.4	59.2	52.3
SRBCT	**DFG**	1	**{GENE2125, GENE2142, GENE2144, GENE2190, GENE2197}**	**66**	**63.8**	**68**	**84.6**	{GENE1802, GENE1820}	68	63.8	66	69.2
	CFG	1	{GENE1113, GENE1114, GENE1115, GENE1679, GENE1680}	60	63.8	38	69.2	**{GENE971, GENE990, GENE1709, GENE1711}**	**69**	**68.4**	**68**	**69.2**
	IGFG	1	{GENE1007, GENE1008, GENE1009, GENE2305, GENE2306}	62	68.4	68	56.1	{GENE773, GENE774, GENE842, GENE843, GENE844}	64	38.4	56	61.5
Lymphoma	**DFG**	1	**{GENE654X, GENE627X, GENE659X}**	**81.6**	**76.9**	**93.8**	**84.6**	{GENE3142X, GENE3096X}	65.5	76.9	77.5	84.6
		2	{GENE579X, GENE585X}	81.6	76.9	93.8	61.5	**{GENE770X, GENE761X, GENE507X}**	**83.6**	**84.6**	**81.6**	**92.3**
	CFG	1	{GENE2559X, GENE1194X}	65.3	76.9	67.3	61.5	{GENE58X, GENE1094X, GENE40X}	65.3	76.9	57.1	69.2
	IGFG	1	{GENE2740X, GENE2741X}	65.3	76.9	59.1	69.2	{GENE3145X, GENE3105X}	65.3	76.9	57.1	61.5

Leukemia	DFG	1	**{M24748_CDS2_S_AT, M31211_S_AT}**	**85.9**	**86.6**	**78.9**	**86.6**	{M32639_AT, M34192_AT}	61.4	86.6	66.6	73.3
		2	**{M96740_AT, S46622_AT, S69232_AT}**	**85.9**	**86.6**	**73.6**	**86.6**	**{X04391_AT, X04707_AT}**	**70.1**	**86.6**	**73.6**	**80**
	CFG	1	{M19722_AT, M19961_AT}	59.6	86.6	63.1	86.6	{AFFX-PHEX-M_AT, AFFX-HUMGAPDH/M33197_3_AT}	59.6	86.6	63.1	73.3
		2	**{U40622_AT, U40714_AT}**	**77.8**	**86.6**	**66.6**	**86.6**	{X04327_AT, X07948_AT, X12433_AT}	61.4	86.6	56.1	73.3
	IGFG	1	{D87002_CDS2_AT,HG2846-HT2983_AT}	59.6	86.6	54.3	73.3	{U61263_AT, U62531_AT, U63717_AT}	63.1	86.6	57.8	80

Table 3: Representative gene subsets obtained by LFSAIRS2 and LFSPAIRS2 based on Associative feature groups

			LFSAIRS2	SVM		KNN		LFSPAIRS2	SVM		KNN	
Data Set	**Group Type**	**No**	**Optimal Gene Subsets**	**10 CV % on**		**10 CV % on**		**Optimal Gene Subsets**	**10 CV % on**		**10 CV % on**	
				Train Set	**Test Set**	**Train Set**	**Test Set**		**Train Set**	**Test Set**	**Train Set**	**Test Set**
Colon	**DFG**	1	{T99498, X15183}	69.3	69.2	67.3	53.8	{U37673,H51015,R54422, R46502}	69.3	61.5	73.6	61.5
		2	{T63539,T93284,U29607}	69.3	69.2	65.3	69.2	{L08069, R93337, T89175}	67.3	61.5	65.3	69.2
	CFG	1	{U25265, T47584, R39130}	63.2	69.2	48.9	61.5	{R21901, R39531, T92736, X14830}	63.2	69.2	59.1	46.1
		2	{M37510, L13738}	63.2	69.2	65.3	46.1	{V00523, T72889}	63.2	69.2	48.9	84.6
	IGFG	1	{D00763,X66839}	63.2	69.2	57.1	69.2	{R09479, M73481,H79349, T41207, T54364, R56052}	63.2	69.2	59.1	69.2
Lungstd	**DFG**	1	**{34171_AT}**	**91.7**	**86.1**	**86.3**	**91.6**	{33969_AT, 34028_AT}	89.6	83.3	87.5	83.3
	CFG	1	{39640_AT,41037_AT, 2047_AT}	82	86.1	84.8	91.6	**{31806_AT, 31840_AT}**	**90.3**	**86.1**	**91**	**86.1**
	IGFG	1	{33324_S_AT,33334_AT}	82	86.1	81.3	80.5	{31980_AT, 32010_AT, 32383_AT, 32396_F_AT}	82.7	86.1	80	83.3
		2	{33336_AT, 33337_AT, 33698_AT}	83.4	86.1	88.9	86.1	{40719_AT, 40721_G_AT}	82	88.8	83.4	91.6
Prostate	**DFG**	1	{34718_AT,35172_AT}	76.5	61.9	62.9	47.6	**{37822_AT, 37823_AT, 38156_AT, 38170_AT}**	**90**	**72.3**	**86.4**	**71.4**
		2	{41405_AT, 41439_AT, 41616_AT}	66.6	67.2	61.7	47.6	**{40727_AT, 41016_AT, 41032_AT}**	**88.8**	**76.1**	**83.9**	**71.4**
	CFG	1	{33714_AT, 33809_AT, 34764_AT}	65.4	42.8	58	47.6	{32663_AT, 32735_AT}	53	52.3	66.6	66.6
	IGFG	1	{36768_AT, 36769_AT, 36770_AT}	61.7	57.1	62.9	57.1	{36763_AT, 36764_AT, 36766_AT, 37051_AT, 37054_AT, 37056_AT}	60.4	52.3	60.4	66.6
SRBCT	**DFG**	1	**{GENE403, GENE415, GENE2048, GENE2081}**	**58**	**61.5**	**72**	**61.5**	{GENE234, GENE253, GENE274}	54	53.8	64	69.2
	CFG	1	{GENE1709, GENE1711}	62	58.4	62	66.1	**{GENE1639, GENE1676, GENE1680}**	**61.3**	**58.9**	**63.2**	**72.4**
	IGFG	1	{GENE767, GENE768, GENE769}	62	58.4	62	63	{GENE864, GENE865}	62	54.6	61.3	60.2
Lymphoma	**DFG**	1	**{GENE1764X, GENE3594X}**	**77.5**	**76.9**	**77.5**	**84.6**	**{GENE2226X, GENE2902X}**	**81.6**	**76.9**	**77.5**	**92.3**
		2	**{GENE2368X, GENE2369X, GENE2370X, GENE2109X, GENE2108X}**	**93.8**	**76.9**	**89.7**	**76.9**	**{GENE2774X,GENE704X, GENE699X}**	**83.6**	**76.9**	**81.6**	**76.9**
	CFG	1	{GENE1910X, GENE2060X, GENE330X, GENE235X}	65.3	76.9	63.2	53.8	{GENE1241X, GENE891X}	65.3	76.9	55.1	61.5
		2	{GENE1269X, GENE1194X}	63.2	76.9	69.2	53.8	{GENE2598X,GENE2616X, GENE2010X}	65.3	76.9	61.2	69.2
	IGFG	1	{GENE2648X,GENE2647X, GENE2684X}	65.3	76.9	55.1	61.5	**{GENE526X, GENE527X}**	**71.4**	**76.0**	**85.7**	**84.6**
Leukemia	**DFG**	1	**{M32639_AT, M34192_AT, U40282_AT, U41387_AT}**	**82.4**	**86.6**	**73.3**	**77.1**	**{M72885_RNA1_S_AT, X58528_S_AT}**	**82.4**	**86.6**	**77.1**	**86.6**
	CFG	1	**{M22382_AT, M23533_AT, X17620_AT, X56465_AT}**	**76.6**	**86.6**	**71.9**	**86.6**	**{D50855_S_AT, Y10807_S_AT}**	**74.9**	**86.6**	**76.6**	**86.6**
	IGFG	1	{U83117_AT, X16546_AT, X78712_AT}	59.6	86.6	61.4	73.3	{M17446_S_AT, U50327_S_AT}	61.4	86.6	59.6	73.3

The average classifier accuracy on the training set and test set separately with using support vector machines and k-NN classifier shown in Table 2 and 3. For Colon data set, 2-gene subsets {T63133, T57630} with 73.4% training accuracy has 92.3% prediction accuracy with KNN classifier based DFG for LFSAIRS1. 2-gene subsets {H22579, X59871} with 73.4% training accuracy has 76.9% prediction accuracy with SVM classifier and 79.1% training accuracy has 77.9% prediction accuracy with KNN classifier based CFG for LFSPAIRS1. For Lungstd data set, 3-gene subsets {33052_at, 33078_at, 33087_s_at} with 95.8% training accuracy has 88.8% prediction accuracy with SVM classifier and 96.5% training accuracy has 92.8% prediction accuracy with KNN classifier based DFG for LFSAIRS1. 3-gene subsets {35052_r_at, 35053_at, 35064_at} with 98.6% training accuracy has 92.2% prediction accuracy with SVM classifier and 96.5% training accuracy has 97.2% prediction accuracy with KNN classifier based DFG for LFSPAIRS1. 1-gene subset {34171_at} with 91.7% training accuracy has 86.1% prediction accuracy with SVM classifier and 86.3% training accuracy has 91.6% prediction accuracy with KNN classifier based DFG for LFSAIRS2. 2-gene subset {31806_at, 31840_at} with 90.3% training accuracy has 86.1% prediction accuracy with SVM classifier and 91% training accuracy has 86.1% prediction accuracy with KNN classifier based CFG for LFSPAIRS2. For Prostate data set, 4-gene subsets {37822_at, 37823_at, 38156_at, 38170_at} with 90% training accuracy has 72.3% prediction accuracy with SVM classifier and 86.4% training accuracy has 71.4% prediction accuracy with KNN classifier based DFG for LFSPAIRS2. 3-gene subsets {40727_at, 41016_at, 41032_at} with 88.8% training accuracy has 76.1% prediction accuracy with SVM classifier and 83.9% training accuracy has 71.4% prediction accuracy with KNN classifier based DFG for LFSPAIRS2. For SRBCT data set, 5-gene subsets {GENE2125, GENE2142, GENE2144, GENE2190, GENE2197} with 66% training accuracy has 63.8% prediction accuracy with SVM classifier and 68% training accuracy has 84.6% prediction accuracy with KNN classifier based DFG for LFSAIRS1. 4-gene subsets {GENE971, GENE990, GENE1709, GENE1711} with 69% training accuracy has 68.4% prediction accuracy with SVM classifier and 68% training accuracy has 69.2% prediction accuracy with KNN classifier based CFG for LFSPAIRS1. 4-gene subsets {GENE403, GENE415, GENE2048, GENE2081} with 58% training accuracy has 61.5% prediction accuracy with SVM classifier and 72% training accuracy has %61.5 prediction accuracy with KNN classifier based DFG for LFSAIRS2. 3-gene subsets {GENE1639, GENE1676, GENE1680} with 61.3% training accuracy has 58.9% prediction accuracy with SVM classifier and 63.2% training accuracy has 72.4% prediction accuracy with KNN classifier based CFG for LFSPAIRS2. For Lymphoma data set, 3-gene subsets {GENE654X, GENE627X, GENE659X} with 81.6% training accuracy has 76.9% prediction accuracy with SVM classifier and 93.8% training accuracy has 84.6% prediction accuracy with KNN classifier based DFG for LFSAIRS1. 3-gene subsets {GENE770X, GENE761X, GENE507X} with 83.6% training accuracy has 84.6% prediction accuracy with SVM classifier and 81.6% training accuracy has 92.3% prediction accuracy with KNN classifier based DFG for LFSPAIRS1. 2-gene subsets {GENE1764X, GENE3594X} with 77.5% training accuracy has 76.9% prediction accuracy with SVM classifier and 77.5% training accuracy has 84.6% prediction accuracy with KNN classifier based DFG for LFSAIRS2. 5-gene subsets {GENE2368X, GENE2369X, GENE2370X, GENE2109X, GENE2108X} with 93.8% training accuracy has 76.9% prediction accuracy with SVM classifier and 89.7% training accuracy has 76.9% prediction accuracy with KNN classifier based DFG for LFSAIRS2. 2-gene subsets {GENE2226X, GENE2902X} with 81.6% training accuracy has 76.9% prediction accuracy with SVM classifier and 77.5% training accuracy has 92.3% prediction accuracy with KNN classifier based DFG for LFSPAIRS2. 3-gene subsets {GENE2774X,GENE704X, GENE699X}with 83.6% training accuracy has 76.9% prediction accuracy with SVM classifier and 81.6% training accuracy has 76.9% prediction accuracy with KNN classifier based DFG for LFSPAIRS2. 2-gene subsets {GENE526X, GENE527X} with 71.4% training accuracy has 76% prediction accuracy with SVM classifier and 85.7% training accuracy has 84.6% prediction accuracy with KNN classifier based IGFG for LFSPAIRS2. For Leukemia data set, 2-gene subsets {M24748_cds2_s_at, M31211_s_at} with 85.9% training accuracy has 86.6% prediction accuracy with SVM classifier and 78.9% training accuracy has 86.6%prediction accuracy with KNN classifier based DFG for LFSAIRS1. 3-gene subsets {M96740_at, S46622_at, S69232_at} with 85.9% training accuracy has 86.6% prediction accuracy with SVM classifier and 73.6% training accuracy has 86.6% prediction accuracy with KNN classifier based DFG for LFSAIRS1. 2-gene subsets {U40622_at, U40714_at} with 77.8% training accuracy has 86.6% prediction accuracy with SVM classifier and 66.6% training accuracy has 86.6% prediction accuracy with KNN classifier based CFG for LFSAIRS1. 2-gene subsets {X04391_at, X04707_at} with 70.1% training accuracy has 86.6% prediction accuracy with SVM classifier and 73.6% training accuracy has 80% prediction accuracy with KNN classifier based DFG for LFSPAIRS1. 4-gene subsets {M32639_at, M34192_at, U40282_at, U41387_at} with 82.4% training accuracy has 86.6% prediction accuracy with SVM classifier and 73.3% training accuracy has 77.1% prediction accuracy with KNN classifier based DFG for LFSAIRS2. 4-gene subsets {M22382_at, M23533_at, X17620_at, X56465_at} with 76.6% training accuracy has 86.6% prediction accuracy with SVM classifier and 71.9% training accuracy has 86.6% prediction accuracy with KNN classifier based CFG for LFSAIRS2. 2-gene subsets {M72885_rna1_s_at, X58528_s_at} with 82.4% training accuracy has 86.6% prediction accuracy with SVM classifier and 77.1% training accuracy has 86.6% prediction accuracy with KNN classifier based DFG for LFSPAIRS2. 2-gene subsets {D50855_s_at, Y10807_s_at} with 74.9% training accuracy has 86.6% prediction accuracy with SVM classifier and 76.6% training accuracy has 86.6% prediction accuracy with KNN classifier based CFG for LFSPAIRS2.

Conclusion

Conclusion

In this paper, we proposed an ensemble gene selection framework to select informative gene subsets. The informative gene subsets obtained from the different type of the associative feature groups mine many tumor-related genes. Based on the obtained optimal gene subsets, we aim to find reliable accuracy on the training set and test set separately. The significance of a gene subset is measured by its frequency occurrence. Each type of the associative feature groups obtained by group-based learning was presented as a candidate solution to Artificial Immune Recognition Systems in order to improve with its meta-dynamics. The presented framework makes it possible to obtain more robust tumor-related genes. The prediction accuracies obtained by SVM and KNN classifiers. The classifier results obtained were compared with six commonly used microarray data sets.

References

He Z., Yu W., (2010), *Stable feature selection for biomarker discovery*. Pubmed.

Loscalzo S , YU L., Ding C. (2009), *Consensus group stable feature selection*, June28–July1, Paris,France.

Brownlee J. (2005). *Artificial immune recognition system (airs): A review and analysis*, Technical Report.

Wang K., Chen K.,Adrian A., (2014), *An improved artificial immune recognition system with the opposite sign test for feature selection* , Knowledge-Based Systems ,Taiwan.

Loscalzo S , YU L., Ding C. (2008), *Stable feature selection via feature groups*, August 24–27, Las Vegas, Nevada, USA.

http://www.cs.waikato.ac.nz/ml/weka/

Y.Saeys, I.Inza, and P. Larranaga, (2008). *A review of feature selection techniques in bioinformatics*, Bioinformatics,vol. 23, no. 19, pp. 392-403.

T. Abeel, T.Helleputte, Y.V. de Peer, P. Dubont, and Y. Saeys, *Robust biomarker identification for cancer diagnosis with ensemble feature selection methods*, Bioinformatics, in press.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, (2002). *Gene selection for cancer classification using support vector machines*, Machine learning, vol. 46, no. 1, pp. 389-422.

J. Dutkowski and A. Gambin, (2007). *On consensus biomarker selection*, BMC Bioinformatics, vol.8, no. Suppl 5, pp. S5,

S. Ma, J. Huang, (2008), *Penalized feature selection and classification in bioinformatics*, Brief. Bioinform. Vol.9 no. 5, pp. 392-403.

Q. Song, J. Ni, G. Wang, (2013). *A fast clustering-based feature subset selection algorithm for high dimensional data*, IEEE Trans. Knowl. Data Eng. Vol. 25 no.1, pp. 1-4.

Oh S. , Lee J.S., Moon B.R. (2004). *Hybrid genetic algorithms for feature selection* .Vol. 26, No.11, November.

Timmis J. and Neal J.,(2000), *Investigating the evolution and stability of a resource limited artificial immune system*, Special Workshop on Artificial Immune Systems, Genetic and Evolutionary Computation Conference (GECCO) 2000, Las Vegas, Nevada, U.S.A., pp. 40-41.

Mark A. Hall, Lloyd A. Smith., (1998) *Practical feature subset selection for machine learning*, In C. Mcdonald (ed.), Computer Science '98 Proceedings of the 21st Australian Computer Science Conference ACSC'98, Perth, 4-6 February, 1998 (pp. 181-191), Berlin: Springer.

Carter J.H., (2000), *The immune system as a model for classification and pattern recognition*, Journal of the American Informatics Association, vol.7.

Watkins A. and Timmis J.,(2002), *Artificial immune recognition system (airs): Revisions and refinements*, 1st International Conference on Artificial Immune Systems (ICARIS2002), University of Kent at Canterbury, pp. 173-181.